digital | recht

Schriften zum Immaterialgüter-, IT-,
Medien-, Daten- und Wettbewerbsrecht

Andrea Linhart

Information aus der Blackbox

Zum Verhältnis von Transparenz und Geheimnisschutz am Beispiel Künstlicher Neuronaler Netze

Band 13

Andrea Linhart

Information aus der Blackbox

Zum Verhältnis von Transparenz und Geheimnisschutz am Beispiel Künstlicher Neuronaler Netze

digital | recht
Schriften zum Immaterialgüter-, IT-, Medien-, Daten- und Wettbewerbsrecht

Herausgegeben von Prof. Dr. Maximilian Becker, Prof. Dr. Katharina de la Durantaye, Prof. Dr. Franz Hofmann, Prof. Dr. Ruth Janal, Prof. Dr. Anne Lauber-Rönsberg, Prof. Dr. Benjamin Raue, Prof. Dr. Herbert Zech

Band 13

Andrea Linhart, geboren 1986, Studium der Romanistik und Rechtswissenschaften in Mainz, Paris, Genf und Berlin; Referendariat in Neuruppin; Rechtsanwältin in Berlin; seit 2020 Referentin im Bundesministerium der Justiz. Das Buch gibt die persönliche Meinung der Verfasserin wieder.
ORCID: 0009-0003-0723-6614

Zugl.: Berlin, Humboldt-Universität zu Berlin, Juristische Fakultät, Dissertation, 2023, u. d. T. „Information Künstlicher Neuronaler Netze zwischen Geheimnisschutz und Transparenz“.

Bibliografische Information der Deutschen Nationalbibliothek:
Die Deutsche Nationalbibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliografie; detaillierte bibliografische Angaben sind im Internet über http://dnb.d-nb.de abrufbar.
Dieses Buch steht gleichzeitig als elektronische Version über die Webseite der Schriftenreihe: http://digitalrecht-z.uni-trier.de/ zur Verfügung.

ISBN: 9783758408281
URN: urn:nbn:de:hbz:385-2023092508
DOI: https://doi.org/10.25353/ubtr-19bc-20e1-b58a

Die Schriftenreihe wird gefördert von der Universität Trier und dem Institut für Recht und Digitalisierung Trier (IRDT).
Anschrift der Herausgeber: Universitätsring 15, 54296 Trier.

UNIVERSITÄT TRIER

IRDT Institut für Recht und Digitalisierung Trier

Meiner Familie

Vorwort

Diese Arbeit wurde im Sommer 2023 von der Juristischen Fakultät der Humboldt-Universität zu Berlin als Dissertation angenommen und kurz darauf verteidigt. Inhalt und Nachweise geben den Stand der Einreichung im Frühjahr 2023 wieder.

Das Verfassen dieser Arbeit war für mich eine große Bereicherung. Mein besonderer Dank gilt daher meinem Doktorvater, Prof. Dr. Herbert Zech, für seine ermutigende Unterstützung und die zahlreichen fachlichen Anregungen bei der Entwicklung meiner Arbeit. Ebenso möchte ich Prof. Dr. Axel Metzger danken für die äußerst zügige Erstellung des Zweitgutachtens sowie PD Dr. Andreas Sattler für die Übernahme des Vorsitzes bei der Disputation.

Herzlich danken möchte ich außerdem Pablo Schumacher und Dr. Florian Kaiser für ihren unermüdlichen fachlichen Rat, ihre Erläuterungen, gerade auch informationstechnischer Grundlagen, sowie für ihre Durchsicht des Manuskripts. Auch viele andere Freunde haben mich zu meiner Forschung beraten und durch ihr Interesse an meiner Arbeit ermutigt, wofür ich ihnen danken möchte.

Der größte Dank gebührt schließlich meiner Familie und aus ihr im Besonderen: meinem Bruder Dr. Friedrich K. E. Linhart, der mich zur Forschung im Bereich Immaterialgüterrechte und *Machine Learning* inspiriert hat und ohne den die Arbeit in dieser Form nicht entstanden wäre. Meinem Freund Alexander Knopf, der mich durch seinen unbeirrbaren Glauben in meine Fähigkeiten motiviert und gestärkt hat. Und natürlich meinen Eltern Ingeborg und Dr. Friedrich Linhart, die – wie in allen Lebenslagen – durch ihr Interesse und ihr Vertrauen meine Promotion begleitet haben. Meiner Familie ist daher diese Arbeit gewidmet.

Berlin, im Juli 2023 Andrea Linhart

Inhaltsverzeichnis

Abbildungsverzeichnis

Abkürzungsverzeichnis

Das folgende Verzeichnis führt nur ungebräuchliche Abkürzungen auf. Für die übrigen verwendeten Abkürzungen wird verwiesen auf *Duden*, Die deutsche Rechtschreibung, Berlin 2020 und *Kirchner* Abkürzungsverzeichnis der Rechtssprache Berlin/Boston 2021.

API	Application Programming Interface
ADM	Algorithmic Decision Making
GOFAI	Good Old Fashioned Artificial Intelligence
IoT	Internet of Things
KI	Künstliche Intelligenz
KI-Haftungs-RL-E	Vorschlag für eine Richtlinie des Europäischen Parlaments und des Rates zur Anpassung der Vorschriften über außervertragliche zivilrechtliche Haftung an künstliche Intelligenz (Richtlinie über KI-Haftung), COM (2022) 496 final
KI-System	System Künstlicher Intelligenz
KI-VO-E	Vorschlag für eine Verordnung des Europäischen Parlaments und des Rates zur Festlegung harmonisierter Vorschriften für Künstliche Intelligenz (Gesetz über Künstliche Intelligenz) und zur Änderung bestimmter Rechtsakte der Union. Zitiert wird die Allgemeine Ausrichtung vom 6.12.2022, 15698/22
KNN	Künstliches Neuronales Netz
LIME	Local Interpretable Model-agnostic Explanations
ML	Machine Learning
MLaaS	Machine-Learning-as-a-service
SHAP	SHapley Additive exPlanations
XAI	Explainable Artificial Intelligence

Al abrir los ojos, vi el Aleph. [...] el lugar donde están, sin confundirse, todos los lugares del orbe, vistos desde todos los ángulos.

Jorge Luis Borges, El Aleph

Einleitung

Eine verbreitete Definition Künstlicher Intelligenz (KI) lautet:

> "*Artificial Intelligence* (A.I.) is the study of how to make computers do things at which, at the moment, people are better."[1]

Es liegt daher in der Natur der Sache, dass Systeme Künstlicher Intelligenz in ihren Fähigkeiten und ihrer Wirkungsweise kontinuierlich zu menschlichem Können aufschließen. Besonders deutlich zeigt sich das in den letzten Jahren im Bereich Künstlicher Neuronaler Netze (KNN). Wenig verwunderlich ist es dann auch, dass diese Annäherung ein gewisses „Unbehagen" beim Menschen auslösen kann.[2] Ein wesentlicher Grund für dieses Unbehagen ist die regelmäßige Intransparenz des Entscheidungsmechanismus eines KI-Systems.

Ihre Folge sind Bestrebungen in verschiedenen Rechtsordnungen, Systeme Künstlicher Intelligenz zu regulieren. Vorreiter ist hier die Europäische Union (EU), die mit ihrem Entwurf einer Verordnung zur Festlegung harmonisierter Vorschriften für Künstliche Intelligenz (KI-VO-E)[3] den ersten Schritt in Richtung einer umfassenden, sektorübergreifenden KI-Regulierung gegangen ist. Eine wichtige Rolle nimmt dabei auch die Transparenz von Systemen Künstlicher Intelligenz ein.

Doch Transparenz stößt im Rahmen von Künstlicher Intelligenz an zwei Grenzen: eine technische und eine rechtliche. Rein faktisch kann sich die verständliche Erklärung der Funktionsweise eines KI-Systems als schwierig, wenn nicht sogar als unmöglich herausstellen. Aus immaterialgüterrechtlicher Sicht stellt

---

[1] „Künstliche Intelligenz ist die Erforschung der Frage, wie man Computer dazu bringen kann, Dinge zu tun, die Menschen zur Zeit besser können." *Rich*, Artificial intelligence, S. 1; Übersetzung durch die Verfasserin.

[2] Siehe *Martini*, Blackbox Algorithmus – Grundfragen einer Regulierung Künstlicher Intelligenz, S. 8.

[3] Vorschlag für eine Verordnung des Europäischen Parlaments und des Rates zur Festlegung harmonisierter Vorschriften für Künstliche Intelligenz (Gesetz über Künstliche Intelligenz) und zur Änderung bestimmter Rechtsakte der Union. Zitiert wird die Allgemeine Ausrichtung vom 6.12.2022, 15698/22.

sich die Frage nach dem Schutz der offengelegten Information, der im Bereich Künstlicher Intelligenz überwiegend in Form von Geschäftsgeheimnissen ausgestaltet ist.
Die vorliegende Arbeit ergründet das Spannungsfeld zwischen Transparenz und Geheimnisschutz mit Blick auf Künstliche Neuronale Netze. Der besondere Fokus liegt dabei auf den Abhängigkeiten zwischen technischen Darstellungsmöglichkeiten von Information, ihrer Verständlichkeit und ihrem Schutz als Geschäftsgeheimnis.

A. Gegenstand der Arbeit

Gegenstand der Arbeit ist die Information trainierter künstlicher neuronaler Netze und ihre Schutzfähigkeit als Geschäftsgeheimnis trotz zunehmender Transparenzpflichten.
Künstliche Neuronale Netze gehören zu den gängigsten und vielseitigsten Techniken im Bereich des sog. *Machine Learning* (Maschinenlernen, *ML*), weshalb sie als technischer Untersuchungsgegenstand für diese Arbeit ausgewählt wurden. Neben ihrer weiten Verbreitung und vielseitigen Einsatzmöglichkeiten besitzen sie jedoch eine weitere Eigenschaft, die ihre Untersuchung besonders interessant macht: Künstliche Neuronale Netze sind häufig ausgesprochen komplex und ihre Entscheidungsfindung daher kaum nachvollziehbar. Regelmäßig werden sie daher als Blackbox bezeichnet.[4]
Dieser Blackbox-Charakter hat Auswirkungen auf unterschiedliche Disziplinen, die sich mit KNN befassen. Im Recht, in der Politik und in den Sozialwissenschaften entspringen der Intransparenz Forderungen nach Transparenz und Erklärbarkeit.[5] Durch Transparenz soll die Entscheidung des KNN nachvoll-

[4] Martini spricht sogar von einer „doppelten Blackbox“: die Funktionsweise eines Algorithmus oder eines Machine Learning-Systems sei opak, darüber hinaus wüssten viele Verbraucher gar nicht, dass sie es mit einem solchen System zu tun haben: *Martini*, Blackbox Algorithmus – Grundfragen einer Regulierung Künstlicher Intelligenz, S. 28 f.

[5] Siehe statt vieler *Martini*, Blackbox Algorithmus – Grundfragen einer Regulierung Künstlicher Intelligenz; *Baase*, A gift of fire; *Wischmeyer*, AöR 2018, 1; *Deutscher Bundestag*, Bericht der Enquete-Kommission Künstliche Intelligenz – Gesellschaftliche Verantwortung und wirtschaftliche, soziale und ökologische Potenziale, BT-Drs. 19/23700; *Bundesregierung*, Strategie Künstliche Intelligenz der Bundesregierung; *Europäische Kommission*, Weißbuch zur Künstlichen Intelligenz.

ziehbar und damit überprüfbar werden.[6] Die Forderungen richten sich an die Informationstechnik, die jedoch ihrerseits auch ein intrinsisches Interesse an der Erklärbarkeit ihrer Methoden hat. Aus diesem Interesse ist ein eigener Forschungsbereich entstanden, die sog. *Explainable Artificial Intelligence* (Erklärbare Künstliche Intelligenz, *XAI*).

Als Ausgangspunkt für die Frage, wie entsprechenden – bestehenden und zukünftigen – Transparenzpflichten unter Wahrung immaterialgüterrechtlicher Interessen nachgekommen werden kann, dienen dieser Untersuchung zwei Transparenzregime. Einerseits die Artikel 13 bis 15 der Datenschutzgrundverordnung (DSGVO) mit ihren Informations- und Auskunftsrechten hinsichtlich der involvierten Logik einer automatisierten Entscheidungsfindung. Andererseits der Entwurf der KI-Verordnung, der an verschiedenen Stellen umfangreiche Dokumentations- und Transparenzpflichten gegenüber unterschiedlichen Empfängergruppen vorsieht.

Es bestehen sehr unterschiedliche Möglichkeiten, die Information eines KNN darzustellen. Die vorliegende Arbeit untersucht zunächst im Detail, was der Informationsgehalt eines Künstlichen Neuronalen Netzes ist. Diese *semantische Information*[7] des KNN bildet den Anknüpfungspunkt für den Schutz als Geschäftsgeheimnis. Darauf aufbauend wird eine Stufenordnung der verschiedenen Darstellungsmöglichkeiten dieser Information entworfen, die von der genauesten, jedoch für den Menschen unverständlichen Darstellung als Maschinencode bis zur Darstellung anhand von Techniken der *XAI* führt. Dabei wird für jede der vier Stufen *syntaktischer Information* untersucht, wieviel möglicherweise geheime Information durch die Darstellungsform offengelegt wird und daran anschließend auch, für welche Empfängergruppe sie sich eignet. Außerdem wird untersucht, welche weitergehenden immaterialgüterrechtlichen Schutzmöglichkeiten für die jeweilige Darstellungsform bestehen.

Ziel der Arbeit ist es zu zeigen, dass ein ausdifferenziertes Stufensystem der Informationsdarstellung Möglichkeiten schafft, das Spannungsverhältnis zwischen Transparenz und Geheimnisschutz im Bereich Künstlicher Neuronaler

---

[6] Das zeigt sich in vielen Beschreibungen der Problemstellung, etwa bei *Lapuschkin u. a.*, Nature Communications 2019, 1 (3); *Voosen*, How AI detectives are cracking open the black box of deep learning, Science, http://www.sciencemag.org/news/2017/07/how-ai-detectives-are-cracking-open-black-box-deep-learning (zuletzt abgerufen am 26.10.2023).; *Hacker/Krestel/Grundmann/Naumann*, Artificial Intelligence and Law 2020, 415 (416).

[7] Siehe zur Unterscheidung von semantischer, syntaktischer und struktureller Information *Zech*, Information als Schutzgegenstand, S. 37 ff.

Netze aufzulösen. Es entsteht ein *System abgestufter Transparenz*, das den Interessen der Geheimnisinhaber ebenso Rechnung trägt wie denen der Transparenzsuchenden und Informationsempfänger.
Die detaillierte Untersuchung der Informationsebenen eines KNN kann zudem weitere Forschung im Bereich des Immaterialgüterrechts unterstützen. Denn ein System künstlicher Intelligenz kommt nur als Schutzobjekt in Frage, wenn der Schutzgegenstand klar abgrenzbar ist und seine Bestandteile definierbar sind.

B. Forschungsstand

Sowohl die Regulierung Künstlicher Intelligenz als auch ihre Schutzfähigkeit sind Gegenstand umfangreicher Forschung.
Im Bereich von Regulierung und Transparenz im Allgemeinen sticht *Martini* mit seiner „Suche nach einem kategorischen Imperativ für die Welt algorithmischer Entscheidungsfindung und maschinellen Lernens“ heraus.[8] Speziell in Bezug auf Transparenz und Erklärbarkeit algorithmischer Entscheidungssysteme existiert – überwiegend in Bezug auf die DSGVO – Forschung in beachtlichem Umfang.[9] Es fehlt jedoch eine differenzierte Analyse, welche Information auf welche Weise offengelegt werden muss, wenn ein KNN im Rahmen einer automatisierten Entscheidungsfindung eingesetzt wird. Einen Anhaltspunkt gibt hier die höchstrichterliche Rechtsprechung zum Scorewert der SCHUFA, die jedoch noch zum alten Recht ergangen ist und sich zudem auf ein Berechnungs-

[8] *Martini*, Blackbox Algorithmus – Grundfragen einer Regulierung Künstlicher Intelligenz, S. VIII; darüber hinaus stützt sich die Untersuchung hier im Wesentlichen auf *Gesellschaft für Informatik*, Technische und rechtliche Betrachtungen algorithmischer Entscheidungsverfahren; *Wischmeyer*, AöR 2018, 1; *Hacker*, NJW 2020, 2142.

[9] *Roßnagel/Nebel/Richter*, ZD 2015, 455; *Wachter/Mittelstadt/Floridi*, IDPL 2017, 76; *Selbst/Powles*, IDPL 2017, 233; *Malgieri/Comandé*, IDPL 2017, 243; *Edwards/Veale*, Duke Law & Technology Review 2017, 18; *Dreyer/Schulz*, Was bringt die Datenschutz-Grundverordnung für automatisierte Entscheidungssysteme?; *Wieder*, in: Taeger, Rechtsfragen digitaler Transformationen - Gestaltung digitaler Veränderungsprozesse durch Recht; *Kaminski*, Berkely Tech L. J. 2019, 190; *Golland*, in: Taeger, Die Macht der Daten und der Algorithmen : Regulierung von IT, IoT und KI; *Kumkar/Roth-Isigkeit*, JZ 2020, 277; *Hacker/Krestel/Grundmann/Naumann*, Artificial Intelligence and Law 2020, 415.

modell bezieht, das sich von einem Künstlichen Neuronalen Netz grundlegend unterscheidet.[10]

Der KI-VO-E wird naturgemäß erst vereinzelt analysiert, eine tiefergehende Untersuchung der Transparenzpflichten wird dabei noch nicht vorgenommen.[11]

Der immaterialgüterrechtliche Schutz von Systemen Künstlicher Intelligenz ist seit einigen Jahren Gegenstand der Forschung, wobei gerade ein möglicher Schutz als Geschäftsgeheimnis eingehend beleuchtet wird.[12] Hier finden sich auch bereits detaillierte Betrachtungen der verschiedenen Bestandteile von KI-Systemen und es können Ansätze einer Differenzierung nach Art der Darstellung der Information ausgemacht werden.[13] Auch der Informationsgehalt Künstlicher Neuronaler Netze wird mitunter für die juristische Analyse herausgearbeitet.[14]

Die Verbindung zwischen Transparenzpflichten und immaterialgüterrechtlichen Schutz von KI wird zwar gezogen, sie lässt jedoch überwiegend die erforderliche Tiefe vermissen.[15]

Gewiss wird in der Forschung zu den Informations- und Auskunftsrechten der DSGVO der Schutz von Geschäftsgeheimnissen mit unterschiedlichen dogmatischen Herleitungen als Begrenzung der Offenlegung anerkannt. Eine konkrete Absteckung des Grenzverlaufs wird jedoch nicht vollzogen. Dies dürfte dem Umstand geschuldet sein, dass das Geschäftsgeheimnis selbst unzureichend bestimmt ist und daher keine Grenzziehung ermöglicht. Auch der KI-VO-E sieht – wie in der Rechtssetzung üblich – pauschal die Wahrung von Geschäftsge-

---

[10] BGH, Urteil v. 28.1.2014, VI ZR 156/13, NJW 2014, 1235.

[11] *Ebers u. a.*, RDi 2021, 528; *Ebert/Spiecker gen. Döhmann*, NVwZ 2021, 1188; *Roos/Weitz*, MMR 2021, 844.

[12] *Surblyté*, in: Ullrich/Hilty/Lamping/Drexl, TRIPS plus 20: From Trade Rules to Market Principles; *Ehinger/Stiemerling*, CR 2018, 761; *Scheja*, CR 2018, 485; *Hartmann/Prinz*, in: Taeger, Rechtsfragen digitaler Transformationen - Gestaltung digitaler Veränderungsprozesse durch Recht; *Sagstetter*, in: Maute/Mackenrodt, Recht als Infrastruktur für Innovation; *Hauck/Cevc*, ZGE 2019, 135; *Söbbing*, CR 2020, 223; *Apel/Kaulartz*, RDi 2020, 24; *Söbbing*, MMR 2021, 111.

[13] Insoweit schon *Springorum*, in: Brunnstein/Sint, Intellectual Property Rights and New Technologies. Proceedings of the KnowRight'95 Conference; insbesondere auch *Ehinger/Stiemerling*, CR 2018, 761; *Söbbing*, CR 2020, 223; *Söbbing*, MMR 2021, 111.

[14] *Zech*, Weizenbaum Series 2020, 1.

[15] Eine Ausnahme bildet hier Alexander, der konkrete Transparenzpflichten für Ranking-Algorithmen aus Sicht des Schutzes von Geschäftsgeheimnissen analysiert: *Alexander*, MMR 2021, 690.

heimnissen vor, ohne jedoch eine Konkretisierung vorzunehmen. In beiden Regelungsregimen und der dazugehörigen Forschung und Rechtsprechung bleibt mithin unklar, welche Informationen eines KNN konkret die Schnittmenge dessen bilden, was einerseits für eine hinlängliche Information des Informationsempfängers erforderlich ist und was andererseits das Geheimhaltungsinteresse des Geheimnisinhabers ausreichend wahrt.
Hier setzt die vorliegende Arbeit an, indem sie kleinteilig herausarbeitet, worin die potentiell als Geschäftsgeheimnis geschützte Information eines KNN besteht und auf welche Weise sie dargestellt werden kann. Steht danach fest, welche Darstellungsformen in welchem Maße geheime Information enthalten, so kann eine Grenzziehung zu Transparenzpflichten jedweder Art erfolgen. Eine gegebenenfalls erforderliche Interessenabwägung kann dann auf einer sicheren tatsächlichen und rechtlichen Grundlage erfolgen.

C. Gang der Untersuchung

Um die Information eines Künstlichen Neuronalen Netzes und die Möglichkeiten ihrer Darstellung bestimmen zu können, werden in einem *ersten Teil* zunächst die informationstechnischen Grundlagen aufgearbeitet. Gegenstand eines *zweiten Teils* sind verschiedene Informationsbegriffe, welche die Analyse der Information eines KNN und ihres immaterialgüterrechtlichen Schutzes erleichtern. Besonderer Fokus liegt hier auf dem Informationsbegriff des Geschäftsgeheimnisgesetzes. Außerdem widmet sich der Teil der Dichotomie von Blackbox und Transparenz und schließt mit der Untersuchung der Transparenzpflichten in DSGVO und KI-VO-E ab.
In einem *dritten Teil* werden die theoretischen Grundlagen aus den vorangegangenen Kapiteln zusammengeführt und auf den Untersuchungsgegenstand, ein trainiertes Künstliches Neuronales Netz, angewendet. Dazu wird zunächst untersucht, was die semantische Information eines KNN ausmacht und unter welchen Voraussetzungen sie immaterialgüterrechtlichen Schutz genießt, insbesondere als Geschäftsgeheimnis. Es folgt die Präsentation der vier Darstellungsstufen der Information eines KNN: Maschinencode (*Erste Darstellungsstufe*), Quellcode und Gewichtsdatei (*Zweite Darstellungsstufe*), Beschreibung, Graph oder Formel (*Dritte Darstellungsstufe*), durch Techniken der *Explainable Artificial Intelligence* (*Vierte Darstellungsstufe*). Hierbei wird jeweils betrachtet, in-

wiefern sich die jeweilige Darstellungsform für die Erfüllung von Transparenzpflichten eignet und inwiefern dies mit der Offenlegung von Geschäftsgeheimnissen verbunden wäre. Für jede Darstellungsform werden anschließend weitergehende bestehende Schutzmöglichkeiten, insbesondere im Urheberrecht, geprüft. Der dritte Teil endet mit einem Exkurs zum *Reverse Engineering* von KNN, das für die Frage des Geheimnisschutzes und einer etwaigen Erforderlichkeit weiterer Schutzrechte von Bedeutung ist.
Der abschließende *vierte Teil* führt die Ergebnisse aus den vorangegangenen Kapiteln zusammen und zeigt, dass sich Geheimnisschutz und Erklärbarkeit im Falle von Künstlichen Neuronalen Netzen keineswegs ausschließen. Sie stehen vielmehr in einer Art symbiotischen Beziehung zueinander, die sich in Form einer abgestuften Transparenz fruchtbar machen lässt.

Teil 1

Informationstechnische Grundlagen

Kapitel 1

Künstliche Neuronale Netze

Nach einer Erläuterung der wichtigsten Begrifflichkeiten im Bereich der Künstlichen Intelligenz widmet sich dieses Kapitel dem Hauptgegenstand dieser Arbeit, den Künstlichen Neuronalen Netzen. Diese werden einer eingehenden Betrachtung unterzogen: ihr Aufbau wird erklärt, sie werden in ihre Bestandteile zerlegt und ihre Funktionsweise wird genau analysiert. Denn Ziel dieses Kapitels ist es, das „technische" Verständnis von Künstlichen Neuronalen Netzen zu vermitteln, das für eine juristische Anknüpfung an das Thema unabdingbar ist.

A. Künstliche Intelligenz und Maschinenlernen

Dieser Abschnitt dient der Klärung grundlegender Begrifflichkeiten im Bereich der Künstlichen Intelligenz. Denn auch wenn Künstliche Intelligenz in unserem alltäglichen Leben und in der gesellschaftlichen Diskussion mittlerweile allgegenwärtig ist, gibt es keine eindeutige Definition des Begriffs. Wer ihm begegnet muss sich daher gezwungenermaßen informieren, was im konkreten Kontext darunter zu verstehen ist. Aufgrund seiner unscharfen Verwendung scheint der Begriff auch für eine wissenschaftliche Untersuchung eher ungeeignet.[16] Dennoch werden hier beispielhaft einige Definitionsversuche dargestellt, um thematisch den Boden für die anschließende Untersuchung zu bereiten.
Anschließend erfolgt eine kurze Erklärung der Methoden des Maschinenlernens, deren wohl wichtigsten Teilbereich die im nächsten Kapitel untersuchten Künstlichen Neuronalen Netze bilden.

[16] Nach Zech eignet sich der Begriff jedoch zur „Kennzeichnung höherer bzw. komplexerer Aufgaben der Informationsverarbeitung", *Zech*, Weizenbaum Series 2020, 1 (10).

Elaine Richs griffige Definitionen Künstlicher Intelligenz ist bereits bekannt:

> "Artificial Intelligence is the study of how to make computers do things at which, at the moment, people are better." [17]

Darin zeigt sich ein Gedanke, der den vielfältigen Definitionen von Künstlicher Intelligenz gemein ist: sie beschreiben Fähigkeiten, die grundsätzlich nur Menschen ausbilden können, und schreiben sie Maschinen zu. Diese Fähigkeiten werden als Ausdruck menschlicher Intelligenz gesehen, weshalb die Grundfrage bei der Beschäftigung mit Künstlicher Intelligenz lautet: Ab wann kann ein System als intelligent beschrieben werden?[18]
Diese Frage wurde im Laufe der Jahrzehnte sehr unterschiedlich beantwortet, am prominentesten wohl durch Alan M. Turing.[19] Der berühmte Mathematiker entwarf bereits 1950 das sogenannte *imitation game*, auch bekannt als Turing-Test. Eine Testperson befindet sich in einem Raum mit zwei Computern, in die sie Fragen eintippen kann. Der eine Computer ist mit einem Menschen verbunden, der andere mit einem Computerprogramm mit dem Namen *Alice*. Wenn die menschliche Testperson aufgrund der Antworten, die ihr in den Computern angezeigt werden, nicht erkennen kann, ob ihr der Mensch oder die Maschine antwortet, kann das antwortende System *Alice* als intelligent beschrieben werden.[20]
In jüngerer Zeit finden sich spezifischere Definitionen von KI auch in Rechtsvorschriften – eine zwingende Konsequenz der Regulierung in diesem Bereich. Hervorzuheben ist hier die im KI-VO-E enthaltene Definition, da die dort vorgesehenen Transparenzpflichten Gegenstand der vorliegenden Untersuchung sein werden.[21]

---

[17] „Künstliche Intelligenz ist die Erforschung der Frage, wie man Computer dazu bringen kann, Dinge zu tun, die Menschen zur Zeit besser können." *Rich*, Artificial intelligence, S. 1; Übersetzung durch die Verfasserin.

[18] Vgl. *Flasiński*, Introduction to Artificial Intelligence, S. 3.

[19] Für einen historischen Überblick der Entwicklung von KI siehe etwa *Flasiński*, Introduction to Artificial Intelligence, S. 3 ff.

[20] Vgl. *Ertel*, Grundkurs Künstliche Intelligenz, S. 4; *Flasiński*, Introduction to Artificial Intelligence, S. 3.

[21] Siehe weitere Definitionen etwa in *Hochrangige Expertengruppe für Künstliche Intelligenz*, Eine Definition der KI: wichtigste Fähigkeiten und Wissenschaftsgebiete, S. 1; *Bundesregierung*, Strategie Künstliche Intelligenz der Bundesregierung, S. 4 f.

In Artikel 3 Nummer 1 KI-VO-E wird ein „„System der künstlichen Intelligenz“ (KI-System)“ definiert als

> „ein System, das so konzipiert ist, dass es mit Elementen der Autonomie arbeitet, und das auf der Grundlage maschineller und/oder vom Menschen erzeugter Daten und Eingaben durch maschinelles Lernen und/oder logik- und wissensgestützte Konzepte ableitet, wie eine Reihe von Zielen erreicht wird, und systemgenerierte Ergebnisse wie Inhalte (generative KI-Systeme), Vorhersagen, Empfehlungen oder Entscheidungen hervorbringt, die das Umfeld beeinflussen, mit dem die KI-Systeme interagieren“.

Das Maschinenlernen ist der wohl bekannteste und auch wirtschaftlich relevanteste Teilbereich der Künstlichen Intelligenz.[22] Es beschäftigt sich mit der Frage, wie Computeralgorithmen aus Erfahrung lernen können und kann wie folgt definiert werden:

> “A computer program is said to learn from experience E with respect to some tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E.” (Von einem Computerprogramm wird gesagt, dass es aus der Erfahrung E in Bezug auf Aufgaben T und das Leistungsmaß P lernt, wenn sich seine Leistung hinsichtlich der Aufgaben T, gemessen durch P, mit der Erfahrung E verbessert.)[23]

Beim Maschinenlernen werden also aus vorhandenen Datensätzen Rückschlüsse für fehlende oder nur latent vorhandene Daten gezogen – die Gesamtheit dieser Rückschlüsse bildet dann ein sogenanntes Modell.[24] Es gibt viele unterschiedliche Möglichkeiten von Modellen, etwa Entscheidungsbäume, Lineare und Logistische Regression, Nächste-Nachbarn-Klassifikation (engl.: k-nearest neighbour), sog. Random Forests und die hier gegenständlichen Künstli-

[22] Siehe dazu nur *Hacker/Krestel/Grundmann/Naumann*, Artificial Intelligence and Law 2020, 415 (416) m.w.N.

[23] *Mitchell*, Machine Learning, S. 2; Übersetzung nach *Hacker*, NJW 2020, 2142 (2142); siehe eingehend auch: *Gesellschaft für Informatik*, Technische und rechtliche Betrachtungen algorithmischer Entscheidungsverfahren, S. 30 ff.; *Zech*, Weizenbaum Series 2020, 1 (27 ff.).

[24] *Ghahramani*, Nature 2015, 452 (452).

chen Neuronalen Netze.[25] Interessant für die vorliegende Arbeit ist nicht so sehr deren Klassifikation, sondern der Umstand, dass manche dieser unterschiedlichen Modelltypen sich mühelos interpretieren lassen, während andere sich einer Interpretation fast völlig verschließen.[26]
Künstliche Neuronale Netze stellen einen sehr relevanten Teilbereich des Maschinenlernens dar, der sich in den letzten Jahren maßgeblich durch den enormen Zuwachs seiner beiden „Treibstoffe" – Daten und Rechnerkapazität – stark entwickelt hat. Ihre Anwendungsfelder und Funktionsweise werden in den nächsten Abschnitten untersucht.

B. Anwendung

Klassischerweise werden in der Informatik Algorithmen entworfen. Darunter werden schrittweise Handlungsanweisungen zur Lösung eines Problems, eingebettet in Software, verstanden.[27] Es gibt jedoch Aufgaben, die sich nicht zufriedenstellend auf diese Weise lösen lassen.[28] Künstliche Neuronale Netze setzen hier an, indem der Lösungsweg nicht durch den Menschen programmiert, sondern durch das Netz anhand von Beispielen „gelernt" wird.[29] Auch für Künstliche Neuronale Netze kommen zwar Algorithmen zum Einsatz. Sie geben jedoch nur das Lernverfahren vor und legen nicht fest, welcher Ausgabewert auf eine konkrete Eingabe folgt. Dies „lernt" das Netz selbsttätig, es programmiert sich insoweit also selbst.[30]
Natürlich stellen Künstliche Neuronales Netz keine aus Erfahrung lernenden Individuen dar. Da es sich jedoch um ein aus der Natur inspiriertes Modell handelt, erfolgt ihre Beschreibung meist mit biologischen und neurowissenschaftlichen Begriffen und die Visualisierung des Modells erfolgt in Anlehnung an die Anordnung von Neuronen im menschlichen Gehirn. Dies darf jedoch nicht darüber hinwegtäuschen, dass es sich schlicht um mathematische Modelle, um eine

[25] Für eine erschöpfende Aufzählung siehe die Einteilung bei *Hacker/Krestel/Grundmann/Naumann*, Artificial Intelligence and Law 2020, 415 (432).
[26] Sog. Whitebox- und Blackbox-Modelle, siehe dazu unten.
[27] Vgl. zum Begriff des Algorithmus *Horowitz/Sahni*, Algorithmen, S. 1; *Martini*, Blackbox Algorithmus – Grundfragen einer Regulierung Künstlicher Intelligenz, S. 17.
[28] Dazu gehört etwa die Bilderkennung, siehe auch *Burrell*, Big Data & Society 2016, 1 (6 f.).
[29] *Lämmel/Cleve*, Künstliche Intelligenz, Wissensverarbeitung - Neuronale Netze, S. 200.
[30] *Knight*, MIT Technology Review 2017, 53 (57).

Nachbildung der biologischen Funktion mit Softwarebausteinen handelt.[31] Bezogen auf das „Lernen" gilt daher:

> „Man sei sich stets bewusst, dass hier mathematisch begründete Adaptionen von Kantengewichten vonstattengehen, die eine gewisse Funktionalität eines Netzes erzeugen. Ein neuronales Netz ist kein handelndes Individuum."[32]

Typische Anwendungsfelder Künstlicher Neuronaler Netze sind die Klassifikation von Daten, die Mustererkennung, die Prognose und die Steuerung.[33] So können die Netze zum Erkennen handschriftlicher Texte und gesprochener Sprache eingesetzt werden (Mustererkennung), zur Bewertung der Kreditwürdigkeit von Bankkunden (Klassifikation), zum Steuern von Robotern und Fahrzeugen, zur medizinischen Diagnostik, zum effizienten Einsatz von Ressourcen, etwa in der Landwirtschaft, zur Komposition musikalischer Werke, zur Verkehrssteuerung, um nur einige Einsatzbereiche zu nennen.[34]

C. Das biologische Vorbild

Im Diskurs über Künstliche Intelligenz tritt die menschliche Neigung, neue Phänomene mit dem menschlichen Dasein zu vergleichen, besonders deutlich zu Tage. Dies zeigt sich schon in den Begriffen „Intelligenz", „Lernen" und „Erfahrung", die im Zusammenhang mit KI gebraucht werden. Nicht immer ist dieser Anthropomorphismus hilfreich in einer von Ängsten vor Kontrollverlust geprägten Debatte über neue Technologien:

> "Of course, talk of neural networks learning from experience should not be taken too seriously. Neural networks do not experience anything.

---

[31] *Lämmel/Cleve*, Künstliche Intelligenz, Wissensverarbeitung - Neuronale Netze, S. 192; Burrell weist daraufhin, wie wenig eingehend sich die rechtswissenschaftliche und sozialwissenschaftliche Kritik mit den mathematischen Hintergründen von Algorithmen beschäftigt: *Burrell*, Big Data & Society 2016, 1 (2).

[32] *Lämmel/Cleve*, Künstliche Intelligenz, Wissensverarbeitung - Neuronale Netze, S. 241.

[33] *Lämmel/Cleve*, Künstliche Intelligenz, Wissensverarbeitung - Neuronale Netze, S. 222.

[34] für Beispiele siehe *Hacker/Krestel/Grundmann/Naumann*, Artificial Intelligence and Law 2020, 415 (419) m.w.N.; *Ertel*, Grundkurs Künstliche Intelligenz, S. 308; *Martini*, Blackbox Algorithmus – Grundfragen einer Regulierung Künstlicher Intelligenz, S. 14 ff.

> They just receive different types of input. But the important point is that they are not fixed in how they respond to inputs. This is because they can change their weights."[35]

Doch der Vergleich mit dem Menschen kann auch sehr fruchtbar sein. Besonders prägnant zeigt sich das im Bereich der Neurowissenschaften und der Forschung zur Verarbeitung von Informationen im menschlichen Gehirn. Dort ermöglichen zwar computerunterstützte (bildgebende) Verfahren immer neue Einblicke in die Funktionsweise des Gehirns, bis hin zur derjenigen eines einzelnen Neurons. Die Verfahren sind jedoch noch insoweit begrenzt, als sie zwar zeigen können, wo Information verarbeitet wird und wie ein einzelnes Neuron auf spezifische Reize reagiert. Sie geben jedoch keinen Aufschluss darüber, wie kognitive Aufgaben tatsächlich ausgeführt werden und so intelligentes Verhalten entstehen kann.[36] Um das herausragende Merkmal des Gehirns, seine Konnektivität, besser zu verstehen, versuchen Wissenschaftler daher, neue mathematische Modelle der Informationsverarbeitung zu entwerfen und so das mögliche Zusammenspiel einer ganzen Schar von Neuronen zu modellieren.[37] Doch KI wird nicht nur modelliert, um die Informationsverarbeitung im menschlichen Gehirn besser zu verstehen, sondern die Informatik selbst bedient sich Erkenntnissen aus den Neuro- und Kognitionswissenschaften, um neue Systeme zu entwerfen.[38] In dieser Wechselbeziehung trägt der Vergleich mit dem menschlichen Organismus insbesondere zur Fortentwicklung eines der vorherrschenden Felder Künstlicher Intelligenz bei.[39]
Die Bezeichnung dieser besonderen Technik des Maschinenlernens als „Neuronale Netze" rührt somit vom Ursprung ihrer Entwicklung her. Das biologische Vorbild kann aber auch helfen, die Funktionsweise besser nachzuvollziehen, weshalb hier dieser Einstieg in die Materie gewählt wird. Dies sollte jedoch nicht

---

[35] *Bermúdez*, Cognitive science: an introduction to the science of the mind, S. 143.

[36] *Bermúdez*, Cognitive science: an introduction to the science of the mind, S. 124; *Ertel*, Grundkurs Künstliche Intelligenz, S. 268.

[37] Vgl. *Bermúdez*, Cognitive science: an introduction to the science of the mind, S. 124.

[38] Siehe dazu eingehend *Mainzer*, Künstliche Intelligenz – Wann übernehmen die Maschinen?, S. 99 ff.

[39] Natürlich werden KNN nicht nur in der „Computational Neuroscience" (theoretische Neurowissenschaft) modelliert, aber ihr Ursprung liegt in diesem Bereich, vgl. *Thomas/Edelman/Crook*, Credit scoring and its applications, S. 57.

darüber hinwegtäuschen, dass die Gemeinsamkeiten zwischen biologischen und künstlichen neuronalen Netzen nur oberflächlich sind.[40]
Im menschlichen Gehirn gibt es etwa hundert Milliarden Neuronen, wobei jedes einzelne mit 1000 bis 10.000 anderen Neuronen verbunden ist, im Hypocampus sogar mit bis zu 50.000.[41] Das Neuron besteht aus einem Zellkörper mit Zellkern und zahlreichen Fortsätzen, den sogenannten Neuriten, die sich wiederum in Dendriten und Axone unterteilen lassen. Jedes Neuron hat mehrere Dendriten, aber nur ein Axon. Das Axon ist dicker als die Dendriten und formt an seinem Ende mehrere Verzweigungen aus, an deren Ende die Synapsen sitzen. Die Dendriten bestehen ihrerseits aus vielen kleineren Verästelungen. Ein Neuron nimmt über die Dendriten elektrische Signale von anderen Neuronen auf. In seinem Zellkörper speichert es die aufgenommene elektrische Spannung, bis diese einen bestimmten Schwellenwert erreicht. Dann „feuert" das Neuron, das heißt es entlädt seine Spannung und gibt sie über einen Spannungsimpuls verteilt an viele andere Neuronen weiter. Dies geschieht über die am Ende des Axons befindlichen Synapsen. Zwischen den Synapsen des feuernden Neurons und den Dendriten eines benachbarten Neurons befindet sich ein kleiner Spalt, der mit den sogenannten Neurotransmittern gefüllt ist. Dabei handelt es sich um chemische Substanzen, die elektrische Spannung von der Synapse des feuernden (präsynaptischen) Neurons an die Dendriten oder den Zellkörper des benachbarten (postsynaptischen) Neurons übertragen können.[42] Ob ein Neuron feuert oder nicht, hängt also entscheidend von den Synapsen ab. Es lassen sich hemmende (inhibitorische) und erregende (excitatorische) Synapsen unterscheiden, erstere hemmen die Weitergabe elektrischer Spannung, letztere fördern sie. Die Höhe der über die hemmenden und erregenden Synapsen eingehenden Spannung ist entscheidend dafür, ob der Schwellenwert des Neurons erreicht wird und es „feuert".[43] Aber die Leitfähigkeit der Synapsen ist nicht gleichbleibend: je mehr Spannungsimpulse sie überträgt, desto leitfähiger wird eine Synapse, keine oder wenige Impulse können sogar zu ihrem Absterben führen.[44]

---

[40] *Russell/Norvig*, Artificial intelligence, S. 750.

[41] *Ertel*, Grundkurs Künstliche Intelligenz, S. 266; *Bermúdez*, Cognitive science: an introduction to the science of the mind, S. 125.

[42] Vgl. zum Ganzen: *Bermúdez*, Cognitive science: an introduction to the science of the mind, S. 125; *Ertel*, Grundkurs Künstliche Intelligenz, S. 266 f.

[43] *Bermúdez*, Cognitive science: an introduction to the science of the mind, S. 126.

[44] *Ertel*, Grundkurs Künstliche Intelligenz, S. 268.

D. Das künstliche Neuron

Das künstliche Neuron greift viele der beschriebenen Merkmale des biologischen Neurons auf. Das erste künstliche Neuron wurde Ende der 1950er Jahre von Frank Rosenblatt entwickelt, der es „Perzeptron" nannte.[45] Es ist eine „Verarbeitungseinheit", bestehend aus Eingabeinformationen, drei mathematischen Funktionen und einem Speicher.[46] Diese Bestandteile werden im Folgenden genauer betrachtet.

Die Terminologie ist bei der Beschreibung Künstlicher Neuronaler Netze nicht immer einheitlich.[47] Hier soll das künstliche Neuron, das gerade den Fokus bildet, mit einem *j* betitelt werden, also etwa das Neuron *j* oder u_j (von engl. *unit*). Eine dem Neuron *j* vorgelagerte Einheit wird dann als Neuron *i* oder u_i bezeichnet.

Die Eingabeinformation entspricht dem Signal, das das biologische Neuron von einem anderen biologischen Neuron erreicht und kann wie in der Natur inhibitorisch oder excitatorisch sein.[48]

Bei der Eingabeinformation handelt es sich um Zahlenwerte $x_0, x_1, ..., x_i$, die den Eingabevektor **X** bilden.[49]

Die Eingaben erreichen das künstliche Neuron entweder unmittelbar aus eingegebenen Daten, oder sie werden ihm von einem anderen künstlichen Neuron im Netz übermittelt. Das hängt davon ab, in welcher Schicht des Netzes sich das empfangende künstliche Neuron befindet.

Liegt das Neuron in der Eingabeschicht, so sind die Eingabewerte $x_0, x_1, ..., x_i$ diejenigen Merkmale, die in die erste Schicht von Neuronen eingegeben werden und den Merkmalsvektor (engl. *feature vector*) **X** bilden.[50] Im Gegensatz zu den

---

[45] Siehe dazu nur *Bermúdez*, Cognitive science: an introduction to the science of the mind, S. 132; *Lämmel/Cleve*, Künstliche Intelligenz, Wissensverarbeitung - Neuronale Netze, S. 203 ff.

[46] *Lämmel/Cleve*, Künstliche Intelligenz, Wissensverarbeitung - Neuronale Netze, S. 192.

[47] Siehe etwa die unterschiedlichen Terminologien einerseits bei *Ertel*, Grundkurs Künstliche Intelligenz, S. 268 f.; *Bermúdez*, Cognitive science: an introduction to the science of the mind, S. 137; und andererseits bei *Lämmel/Cleve*, Künstliche Intelligenz, Wissensverarbeitung - Neuronale Netze, S. 192; *Ehinger/Stiemerling*, CR 2018, 761 (762).

[48] *Bermúdez*, Cognitive science: an introduction to the science of the mind, S. 125.

[49] *Flasiński*, Introduction to Artificial Intelligence, S. 159.

[50] Vgl. *Ertel*, Grundkurs Künstliche Intelligenz, S. 201; *Flasiński*, Introduction to Artificial Intelligence, S. 159 ff.; siehe auch die Definition von Eingabedaten in Art. 3 Nr. 32 KI-VO-E: „die in ein KI-System eingespeisten oder von diesem direkt erfassten Daten, auf deren Grundlage das System ein Ergebnis (Ausgabe) hervorbringt".

Eingabeinformationen, die verborgene oder Ausgabeneuronen erreichen, werden diese Eingabewerte noch nicht gewichtet. Es gibt also keine gerichtete Verbindung zu dieser ersten Schicht von Neuronen.[51]
Solche Eingabewerte oder auch Eingabevariablen können beispielsweise die Merkmale eines Bankkunden sein, der eine Kreditkarte möchte.[52] Setzt eine Bank zur Bewertung der Kreditwürdigkeit des Kunden ein Künstliches Neuronales Netz ein, so muss sie zunächst relevante Merkmale des Kunden bestimmen und die zugehörigen Daten vorbereiten. Dies können zum Beispiel Alter, Geschlecht, Kredithistorie, Wohnort, und viele mehr sein. Diese Merkmale bilden dann die Eingabewerte $x_0, x_1, ..., x_i$, also etwa $x_{Alter}, x_{Geschlecht}, x_{Kredithistorie}, ...$, die, in Zahlenwerte transformiert, in die erste Schicht des neuronalen Netzes eingegeben werden.[53]
Eingabeinformationen, die ein verborgenes Neuron *j* oder ein Ausgabeneuron *j* erreichen, stellen den Aktivierungszustand x_i des vorgeschalteten Neurons *i* dar. Sie sind mit einem numerischen Gewicht w_{ji} (engl. *weight*) versehen, wobei ein Gewicht zwischen -1 und 1 üblich ist. Im biologischen Modell gedacht entspricht dann ein positives Gewicht einer excitatorischen, ein negatives Gewicht einer inhibitorischen Synapse.[54] Das Gewicht wird mit w_{ji} betitelt, da es das Gewicht der Verbindung von Neuron *i* zu Neuron *j* ist.[55]
Das künstliche Neuron erreichen also die mit einem numerischen Gewicht ($w_{j1}, w_{j2}, ..., w_{ji}$) versehenen Eingabewerte $x_0, x_1, ..., x_i$. Die erste Funktion, die diese eintreffende Information verarbeitet, ist die sogenannte Propagierungsfunktion. Durch sie wird aus allen eintreffenden Eingaben $x_1, x_2, ..., x_i$ und ihren Gewichten $w_{j1}, w_{j2}, ..., w_{ji}$ die Netzeingabe net_i berechnet, die die Grundlage für weitere Berechnungen im Neuron *j* ist. Die Netzeingabe ist also die Zusammenfassung aller aus vorgeschalteten Neuronen (oder aus der Umgebung, falls es sich um ein Eingabeneuron handelt) eintreffenden Informationen.[56]

---

[51] *Lämmel/Cleve*, Künstliche Intelligenz, Wissensverarbeitung - Neuronale Netze, S. 197.

[52] *Thomas/Edelman/Crook*, Credit scoring and its applications, S. 57.

[53] Siehe dazu *Lämmel/Cleve*, Künstliche Intelligenz, Wissensverarbeitung - Neuronale Netze, S. 233 ff.

[54] *Bermúdez*, Cognitive science: an introduction to the science of the mind, S. 126.

[55] Das entspricht der Logik der Vektorrechnung, wo die Verbindung von i nach j als wji bezeichnet wird. Siehe zur weiteren Begründung speziell für KNN *Kruse u. a.*, Computational Intelligence, S. 34.

[56] *Lämmel/Cleve*, Künstliche Intelligenz, Wissensverarbeitung - Neuronale Netze, S. 192.

Die Propagierungsfunktion ist die Summe über alle (gewichteten) Eingabeinformationen:[57]

$$f_{propi} = net_i = \sum_{i=1}^{n} w_{ji}\, x_i$$

Zur Erläuterung der Formel: Das große Sigma Σ ist das Summenzeichen. Addiert werden die Produkte von w_{ji} und x_i, wobei i Werte zwischen 1 (daher $i=1$ unter dem Sigma) und n (daher n über dem Sigma) annehmen kann.[58] Erreicht dieser Wert net_i einen gewisse Schwellenwert, so „feuert" das künstliche Neuron und gibt ein Signal an die nachgeschalteten Neuronen weiter.[59]

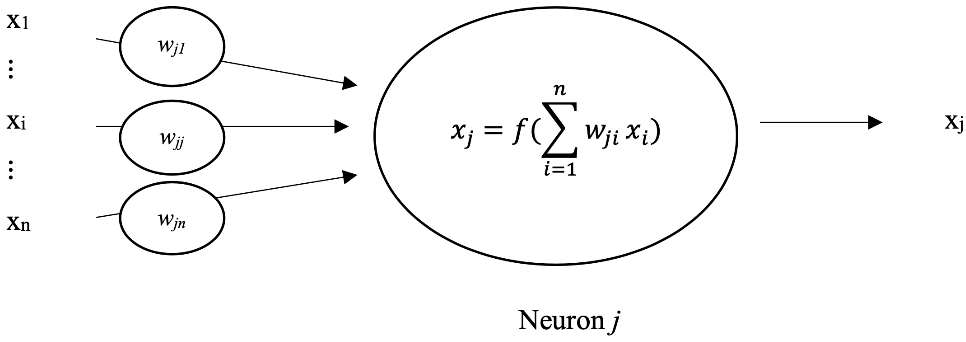

Abb. 1 *Aufbau eines künstlichen Neurons, das auf die Propagierungsfunktion die Aktivierungsfunktion f anwendet, mit Eingabewerten und Gewichten.*[60]

Nachdem berechnet wurde, welcher Zahlenwert das Neuron *j* als Eingabe erreicht, muss noch berechnet werden, welchen Zahlenwert das Neuron *j* seinerseits an die nachgelagerten Neuronen weitergibt.[61] Übertragen auf das biologische Modell würde dies der Stärke des elektrischen Signals entsprechen. Die Berechnung dieser Ausgabe des Neurons *j* an seine Nachbarneuronen erfolgt

[57] *Lämmel/Cleve*, Künstliche Intelligenz, Wissensverarbeitung - Neuronale Netze, S. 193.

[58] Vgl. *Bermúdez*, Cognitive science: an introduction to the science of the mind, S. 126.

[59] *Bermúdez*, Cognitive science: an introduction to the science of the mind, S. 126.

[60] Adaptiert nach *Ertel*, Grundkurs Künstliche Intelligenz, S. 269 und; *Bermúdez*, Cognitive science: an introduction to the science of the mind, S. 126.

[61] *Bermúdez*, Cognitive science: an introduction to the science of the mind, S. 126.

durch eine sogenannte Aktivierungsfunktion.[62] Mit dieser wird also die Aktivierung eines künstlichen Neurons berechnet.
Während als Propagierungsfunktion in der Praxis ausschließlich die soeben beschriebene Formel benutzt wird,[63] stehen für die Aktivierungsfunktion verschiedene Funktionen zur Auswahl, wobei die Wahl der Aktivierungsfunktion einen Einfluss auf das „Verhalten" des Künstlichen Neuronalen Netzes hat.
Typische Aktivierungsfunktionen sind die lineare Funktion, die Schwellwertfunktion und die Sigmoid-Funktion (siehe **Abb. 2**).[64]
Bei der linearen Funktion $f(x) = x$ wächst der Ausgabewert x_j des Neurons j proportional zu seinem Eingabewert net_i. Das Neuron gibt also seine Eingabewerte an die nächsten Neuronen weiter.[65] Problematisch daran ist, dass es keine Begrenzung gibt und die Funktionswerte ins Unermessliche wachsen können, was zu Konvergenzproblemen im Netz führen kann.[66]
Dem kann durch eine Schwellwertfunktion abgeholfen werden. Die Schwellwertfunktion gibt den Wert 0, also kein Signal aus, bis der Eingabewert net_i eine gewisse Schwelle Θ erreicht. Ab dann kann der Ausgabewert entweder proportional zum Eingabewert wachsen. Alternativ kann auch eine binäre Schwellwertfunktion (die sogenannte Heavisidesche Stufenfunktion) eingesetzt werden, bei der die Ausgabewerte nur 0 und 1 sein können. Vor Erreichen des Schwellwert Θ ist das Ergebnis der Aktivierungsfunktion und damit auch die Ausgabe des Neurons 0, nach Erreichen des Schwellwerts Θ ist es 1.[67] Die Heavisidesche Stufenfunktion kommt in binären Neuronen zum Einsatz, also solchen Neuronen, die lediglich die Werte 1 (= ja) oder 0 (= nein) ausgeben können.[68]

---

[62] Vgl. *Ertel*, Grundkurs Künstliche Intelligenz, S. 269; genaugenommen wird die Ausgabe erst aus der Aktivierung durch eine Ausgabefunktion berechnet. Da die Ausgabefunktion jedoch meist die Identität ist, f(x) = x, wird dieser Zwischenschritt hier außer Acht gelassen. Vgl. dazu *Lämmel/Cleve*, Künstliche Intelligenz, Wissensverarbeitung - Neuronale Netze, S. 194; *Flasiński*, Introduction to Artificial Intelligence, S. 161.

[63] *Lämmel/Cleve*, Künstliche Intelligenz, Wissensverarbeitung - Neuronale Netze, S. 193.

[64] Vgl. zu den verschiedenen Möglichkeiten und ihren Implikationen *Lämmel/Cleve*, Künstliche Intelligenz, Wissensverarbeitung - Neuronale Netze, S. 193 f. *Ertel*, Grundkurs Künstliche Intelligenz, S. 269 f. *Bermúdez*, Cognitive science: an introduction to the science of the mind, S. 126 f.

[65] *Ertel*, Grundkurs Künstliche Intelligenz, S. 269.

[66] *Ertel*, Grundkurs Künstliche Intelligenz, S. 269.

[67] Vgl. zum Ganzen *Bermúdez*, Cognitive science: an introduction to the science of the mind, S. 126; *Ertel*, Grundkurs Künstliche Intelligenz, S. 269.

[68] *Ertel*, Grundkurs Künstliche Intelligenz, S. 269.

Formal ausgedrückt lautet die Funktion:

$$H_\Theta(x) = \begin{cases} 0 & falls\ x < \Theta \\ 1 & sonst \end{cases}.$$ [69]

Propagierungsfunktion und Aktivierungsfunktion (hier Heavisidesche Stufenfunktion) zusammengenommen ergibt sich für die Ausgabe x_j des binären Neurons j also folgende Berechnung:

$$x_j = \begin{cases} 0 & falls \sum_{i=1}^{n} w_{ji}\, x_i < \Theta \\ 1 & sonst \end{cases}.$$ [70]

Wie aus **Abb. 2** ersichtlich, ist die Stufenfunktion jedoch unstet. Diese Unstetigkeit kann durch eine sogenannte Sigmoid-Funktion geglättet werden, die dann s-förmig verläuft.[71] Eine sigmoide Funktion, die in kleineren, vorwärts gerichteten KNN häufig zum Einsatz kommt, ist die logistische Funktion:

$$f_{log}(x) = \frac{1}{1 + e^{-c \cdot x}}.$$ [72]

Die logistische Funktion gibt dann Werte zwischen 0 und 1 aus. Sollen Werte zwischen -1 und 1 ausgegeben werden, wird der Tangens Hyperbolicus benutzt ($f_{tan}(x) = \tanh c \cdot x$).[73]

[69] *Ertel*, Grundkurs Künstliche Intelligenz, S. 269.

[70] *Ertel*, Grundkurs Künstliche Intelligenz, S. 269.

[71] Vgl. *Ertel*, Grundkurs Künstliche Intelligenz, S. 270; *Lämmel/Cleve*, Künstliche Intelligenz, Wissensverarbeitung - Neuronale Netze, S. 193.

[72] *Lämmel/Cleve*, Künstliche Intelligenz Wissensverarbeitung - Neuronale Netze, S. 193.

[73] Vgl. *Thomas/Edelman/Crook*, Credit scoring and its applications, S. 58; *Lämmel/Cleve*, Künstliche Intelligenz, Wissensverarbeitung - Neuronale Netze, S. 194.

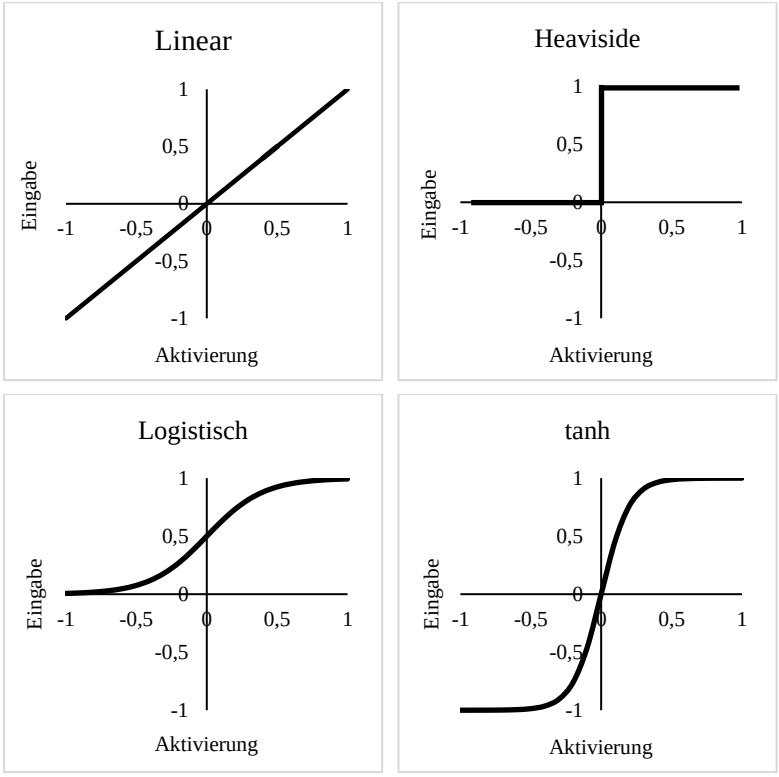

Abb. 2 *Aktivierungsfunktionen*

Bei der sigmoiden Funktion als Aktivierungsfunktion zeigen sich gewisse Parallelen zum biologischen Neuron. Eingaben, die unterhalb eines gewissen Schwellenwerts bleiben, haben kaum einen Einfluss, oberhalb des Schwellenwerts sind Ein- und Ausgabe mehr oder weniger proportional, und bei Eingabewerten weit über dem Schwellenwert steigt die Ausgabe kaum noch an: Die „maximum firing rate" des Neurons ist erreicht. [74]

Aus den vorstehenden Erläuterungen dürfte sich nun Lämmel und Cleves Definition eines künstlichen Neurons erschließen:

> „Ein Neuron ist eine Verarbeitungseinheit, die die über die gewichteten Verbindungen eingehenden Werte geeignet zusammenfasst (Propagierungsfunktion) und daraus mittels einer Aktivierungsfunktion unter

[74] Zum Ganzen *Bermúdez*, Cognitive science: an introduction to the science of the mind, S. 127.

Beachtung eines Schwellwertes einen Aktivierungszustand ermittelt. Aus dieser Aktivierung bestimmt eine Ausgabefunktion die Ausgabe des Neurons."[75]

E. Die Netzarchitektur

Künstliche Neuronale Netzwerke bestehen aus mehreren künstlichen Neuronen, die miteinander verbunden und in Schichten (engl. *layers*) angeordnet werden (vgl. **Abb. 3**).[76] Das Verbindungsgewicht w_{ji} charakterisiert die Verknüpfung zwischen den zwei künstlichen Neuronen i und j.[77] Das einfachste Netz besteht aus einer Eingabeschicht (engl. *input layer*), in die zu verarbeitende Daten eingegeben werden, und einer Ausgabeschicht (engl. *output layer*), die das Ergebnis der Verarbeitung darstellt. Zwischen diesen beiden Schichten können zahlreiche weitere Schichten angebracht werden, die dann als verborgene oder verdeckte Schichten (engl. *hidden layers*) bezeichnet werden.[78]
Mehrlagige Netze können weiter untergliedert werden nach der Art, wie Information innerhalb des Netzes verarbeitet wird. Sehr verbreitet sind die sogenannten vorwärtsgerichteten (engl. *feedforward*) Netze, bei denen die Eingabeinformation durch die Verbindungen von den Neuronen der Eingabeschicht, über die Neuronen der verborgenen Schicht(en) bis zur Ausgabeschicht weitergeben wird. Es sind somit alle Neuronen der verschiedenen Schichten in eine Richtung miteinander verbunden, jedoch nicht die Neuronen einer Schicht untereinander.[79]
Es gibt jedoch auch rückgekoppelte (engl. *recurrent*) Netze, bei denen auch Verbindungen von einer nachgelagerten zur vorgelagerten Schicht gehen.[80] Darüber

---

[75] *Lämmel/Cleve*, Künstliche Intelligenz, Wissensverarbeitung - Neuronale Netze, S. 192.

[76] Da es sich bei Neuronalen Netzen mathematisch um Graphen handelt, werden die Neuronen auch als „Knoten" bezeichnet, die durch „gerichtete und gewichtete Kanten" verbunden sind, *Lämmel/Cleve*, Künstliche Intelligenz, Wissensverarbeitung - Neuronale Netze, S. 309.

[77] *Lämmel/Cleve*, Künstliche Intelligenz, Wissensverarbeitung - Neuronale Netze, S. 196.

[78] Vgl. zum Ganzen *Lämmel/Cleve*, Künstliche Intelligenz, Wissensverarbeitung - Neuronale Netze, S. 197; *Flasiński*, Introduction to Artificial Intelligence, S. 168.

[79] *Bermúdez*, Cognitive science: an introduction to the science of the mind, S. 137; *Lämmel/Cleve*, Künstliche Intelligenz, Wissensverarbeitung - Neuronale Netze, S. 203.

[80] *Flasiński*, Introduction to Artificial Intelligence, S. 170.

hinaus können partiell rückgekoppelte, autoassoziative und wachsende Netzarchitekturen unterschieden werden.[81]
Die Netzarchitektur ergibt sich also aus der Anzahl der künstlichen Neuronen, der Anzahl der Schichten und den Verbindungen zwischen den Neuronen.[82] Die Wahl dieser Parameter ist bestimmt durch die zu lösende Aufgabe und ist – zusammen mit der Wahl des Lernverfahrens – maßgeblich für Einsatzmöglichkeit und Leistungsfähigkeit des Künstlichen Neuronalen Netzes.[83]

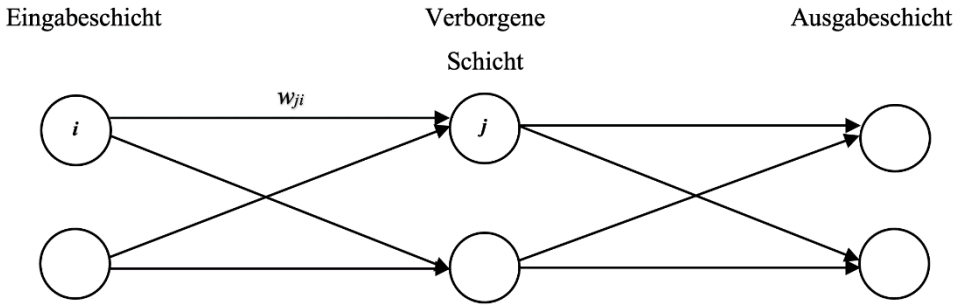

Abb. 3 *Aufbau eines dreilagigen vorwärtsgerichteten Netzes mit sechs Neuronen und gewichteten Verbindungen*

Bei der Implementierung eines Künstlichen Neuronalen Netzes müssen somit verschiedene Aspekte der Netzarchitektur durch den Programmierer bestimmt werden.[84] Dieser Aufgabe kommt einige Bedeutung zu. Denn das Design der Netzarchitektur beeinflusst die Leistungsfähigkeit des Neuronalen Netzes. So hängt etwa die Generalisierungsfähigkeit des Netzes maßgeblich von der Netzarchitektur ab. Der sogenannte Generalisierungsfehler – die Abweichung der Ergebnisse bei der Verarbeitung von Testdaten von den Ergebnissen mit Trainingsdaten – wird mitbestimmt durch die Anzahl der Neuronen, vor allem in den verborgenen Schichten, und die Art und Anzahl der Verbindungen zwischen ihnen.[85]
Die Anzahl der Neuronen hat also einen Einfluss auf die Funktion des Netzes. In der Eingabe- und der Ausgabeschicht ist sie jedoch durch die zu lösende Auf-

[81] *Lämmel/Cleve*, Künstliche Intelligenz, Wissensverarbeitung - Neuronale Netze, S. 200.
[82] Vgl. *Lämmel/Cleve*, Künstliche Intelligenz, Wissensverarbeitung - Neuronale Netze, S. 196.
[83] *Lämmel/Cleve*, Künstliche Intelligenz, Wissensverarbeitung - Neuronale Netze, S. 198, 200.
[84] *Thomas/Edelman/Crook*, Credit scoring and its applications, S. 63.
[85] *Lämmel/Cleve*, Künstliche Intelligenz, Wissensverarbeitung - Neuronale Netze, S. 240.

gabe weitgehend vorgegeben.[86] Das Beispiel der Buchstabenerkennung (siehe Abschnitt G) macht dies deutlich: während die Anzahl der Eingabeneuronen durch die Anzahl der Bildpunkte im gewählten Rasterbild bestimmt wird (hier 5x7), ist die Anzahl der Ausgabeneuronen durch die 26 Großbuchstaben des lateinischen Alphabets vorgegeben. Für ein Netz zur Bewertung der Kreditwürdigkeit wird die Anzahl der Eingabeneuronen durch die Anzahl der Variablen in den Trainingsdaten bestimmt,[87] die Anzahl der Ausgabeneuronen durch die gewünschte Klassifizierung (etwa ein Neuron für „gute" und ein zweites für „schlechte" Kreditwürdigkeit).
Schwieriger ist die Bestimmung der Anzahl der verborgenen Neuronen. In einem vorwärtsgerichteten Netz etwa führt eine innere Schicht mit zu vielen Neuronen schnell zur sogenannten Überanpassung (engl. *overfitting*). Das zu groß konstruierte Netz kann die Trainingsmuster auswendig lernen und ist dementsprechend nicht in der Lage, auf neue Daten zu generalisieren.[88] Die Anzahl zufällig gewählter Eingabemuster, die ein Künstliches Neuronales Netz fehlerlos zufällig gewählten Ausgabemustern zuordnen kann, wird als Speicherkapazität des Netzes bezeichnet.[89] Bis zur Grenze dieser Speicherkapazität, also der Fähigkeit zum „Auswendiglernen", ist der Trainingsfehler des Netzes gleich Null. Wird die Anzahl der verborgenen Neuronen reduziert, steigt der Trainingsfehler solange an, bis das Netz die Muster nicht mehr unterscheiden kann. Durch das mehrfache Training mit Musterdatenmengen kann die ideale Größe der verborgenen Schicht näherungsweise bestimmt werden.[90]
Auch die Anzahl der inneren Schichten muss durch aufwendige Experimente ermittelt werden.[91] Die Anzahl der verdeckten Schichten hat einen Einfluss darauf, wie Merkmale zueinander kombiniert und klassifiziert werden können.[92] Neuronale Netze ohne eine verdeckte Schicht können nur linear separable Klassen abbilden.[93] Das sind solche Klassen, deren Mengen sich durch eine Hyperebene, im

[86] *Lämmel/Cleve*, Künstliche Intelligenz, Wissensverarbeitung - Neuronale Netze, S. 241.
[87] *West*, Computers & Operations Research 2000, 1131 (1142).
[88] *Lämmel/Cleve*, Künstliche Intelligenz, Wissensverarbeitung - Neuronale Netze, S. 241.
[89] *Lämmel/Cleve*, Künstliche Intelligenz, Wissensverarbeitung - Neuronale Netze, S. 243.
[90] *Lämmel/Cleve*, Künstliche Intelligenz, Wissensverarbeitung - Neuronale Netze, S. 241 f.
[91] *Ehinger/Stiemerling*, CR 2018, 761 (764).
[92] Vgl. *Thomas/Edelman/Crook*, Credit scoring and its applications, S. 63.
[93] *Thomas/Edelman/Crook*, Credit scoring and its applications, S. 63; *Ertel*, Grundkurs Künstliche Intelligenz, S. 204.

zweidimensionalen Raum durch eine Gerade, trennen lassen.[94] Sobald eine verdeckte Schicht eingebaut und eine nichtlineare Aktivierungsfunktion benutzt wird, kann das Netz auch nicht-linear separable Klassen abbilden.[95] Bei der Mustererkennung machen komplexere Muster mehrere Schichten erforderlich.[96]
Darüber hinaus haben Art und Anzahl der Gewichte beziehungsweise der Verbindungen zwischen den Neuronen Einfluss auf die Leistungsfähigkeit des Netzes. Zur Optimierung eines Netzes können daher Verbindungen, deren Gewichte im Trainingsprozess stets Werte um die Null annehmen, entfernt oder umverteilt werden.[97] Der Entwickler muss auch die optimale Art der Verbindungen im Netz herausfinden, also ob für die Problemstellung etwa ein rein vorwärtsgerichtetes, ein (partiell) rückgekoppeltes oder ein autoassoziatives Netz am besten geeignet ist.
Die Entwicklung der passenden Netzarchitektur ist somit sehr aufwendig. Hinzu kommt, dass schon kleine Veränderungen der Problemstellung eine Anpassung der Netzarchitektur erforderlich machen können.[98]

F. Lernverfahren

Sind die Netzarchitektur und die Funktionsweise der künstlichen Neuronen festgelegt, muss das Netz „trainiert" werden. Dabei besteht das „Lernen" des Netzes in einer Anpassung der Gewichte zwischen den einzelnen Neuronen, gegebenenfalls auch in einer Anpassung des Schwellwerts:[99]

> „Lernen" ist hier Mathematik pur: Mittels mathematischer Verfahren wird der Netzfehler minimiert."[100]

---

[94] Vgl. *Ertel*, Grundkurs Künstliche Intelligenz, S. 199.

[95] *Thomas/Edelman/Crook*, Credit scoring and its applications, S. 63.

[96] *Ehinger/Stiemerling*, CR 2018, 761 (763).

[97] *Lämmel/Cleve*, Künstliche Intelligenz, Wissensverarbeitung - Neuronale Netze, S. 244.

[98] *Lämmel/Cleve*, Künstliche Intelligenz, Wissensverarbeitung - Neuronale Netze, S. 241.

[99] *Bermúdez*, Cognitive science: an introduction to the science of the mind, S. 132; siehe auch Entwurf der KI-Verordnung, Allgemeine Ausrichtung, Erwägungsgrund 6a: „Der Begriff ‚Lernen' bezeichnet den Rechenvorgang, bei dem anhand von Daten die Parameter eines Modells optimiert werden, das als mathematische Konstruktion auf der Grundlage von Eingabedaten Ergebnisse hervorbringt."

[100] *Lämmel/Cleve*, Künstliche Intelligenz, Wissensverarbeitung - Neuronale Netze, S. 212.

Grundsätzlich werden drei verschiedene Arten des Trainings beziehungsweise des Lernens unterschieden.
Das überwachte Lernen (engl. *supervised learning*) erfolgt anhand von Beispielpaaren aus Eingabedaten und gewünschten Ausgabewerten.[101] Gibt das Netz für einen Eingabewert den falschen Ausgabewert aus, so bekommt es eine Fehlermeldung.[102] Auf diese Weise wird das Netz trainiert, bis es die richtigen Werte ausgibt. Nach diesem Training muss dann die Generalisierungsfähigkeit des Netzes anhand eines Sets von Testdaten überprüft werden, die nicht Teil der Trainingsdaten sind. So wird sichergestellt, dass das Netz die Datensätze nicht schlicht „auswendig gelernt" (sog. *overfitting*), sondern ein Klassifikationsmodell entworfen hat, das auch mit neuen, unbekannten Daten funktioniert.[103]
Das bestärkende Lernen (engl. *reinforcement learning*) kommt zum Einsatz, wenn dem Netz keine klare Rückmeldung darüber gegeben werden kann, wie stark die Abweichung des tatsächlichen vom gewünschten Ausgabewert ist. Das Netz bekommt also nur die grobe Meldung, ob die Ausgabe richtig oder falsch ist.[104] Typisches Anwendungsfeld für das bestärkende Lernen ist die Robotik. Denn die komplexen und vielfältigen Aufgaben, die ein Roboter zu lösen hat, können weder programmiert noch durch ausreichende Daten im Vorhinein trainiert werden.[105]
Das dritte Lernverfahren ist das unüberwachte Lernen (engl. *unsupervised learning*). Es wird eingesetzt, wenn die richtigen Ausgabewerte unbekannt sind, also zum Beispiel bei Klassifizierungen, bei denen die Einteilung der Klassen nicht bekannt ist.[106] Bei dieser Methode muss das Netz dann ohne die Hilfe von Beispielen lernen, welches der richtige Ausgabewert für eine Eingabe ist.[107]
Diese Verfahren beschreiben jedoch nur die Art des Lernens, sie geben keinen Aufschluss darüber, wie dieses Lernen, also die Anpassung der Verbindungsgewichte, vonstatten geht.

---

[101] Vgl. *Lämmel/Cleve*, Künstliche Intelligenz, Wissensverarbeitung - Neuronale Netze, S. 199.
[102] *Bermúdez*, Cognitive science: an introduction to the science of the mind, S. 132.
[103] Vgl. *Lämmel/Cleve*, Künstliche Intelligenz, Wissensverarbeitung - Neuronale Netze, S. 198.
[104] *Lämmel/Cleve*, Künstliche Intelligenz, Wissensverarbeitung - Neuronale Netze, S. 199.
[105] *Ertel*, Grundkurs Künstliche Intelligenz, S. 313.
[106] Sog. Clusterungen, vgl. *Lämmel/Cleve*, Künstliche Intelligenz, Wissensverarbeitung - Neuronale Netze, S. 199.
[107] Vgl. *Lämmel/Cleve*, Künstliche Intelligenz, Wissensverarbeitung - Neuronale Netze, S. 199.

Das Lernen wird durch einen Lernalgorithmus gesteuert. Je nach Art des Lernens (überwacht, bestärkend, unüberwacht) und den Eigenschaften des zu trainierenden Netzes kommen unterschiedliche solche Algorithmen zum Einsatz.
Typische Lernalgorithmen für das überwachte Lernen sind die sogenannte Delta-Regel[108] für einlagige Netze und der Backpropagation-Algorithmus für mehrlagige Netze.[109] Verstärkendes Lernen kann beispielsweise durch das sogenannte Q-Lernen erfolgen.[110] Beim unüberwachten Lernen kommt häufig die sogenannte Hebb-Regel zum Einsatz.[111]
Anhand des hier gewählten Beispiel-Netzes (letters.net) soll mit dem Backpropagation-Algorithmus das meistgenutzte[112] Lernverfahren vorgestellt werden.
Durch den Backpropagation-Algorithmus wird die Abweichung des gewollten vom tatsächlichen Ausgabewert berechnet und dem Netz ein entsprechendes Fehlersignal als Feedback gegeben.[113] Anhand dieses Fehlersignals kann das Netz auch die notwendige Aktivierung x_j eines Neurons j in einer verborgenen Schicht bestimmen, für das es auch beim überwachten Lernen keinen korrekten Beispielwert kennt. Denn dem Netz können nur die korrekten Ausgabewerte für die Ausgabeschicht zur Verfügung gestellt werden. Durch den Backpropagation-Algorithmus berechnet das Netz sozusagen den Beitrag der vorgeschalteten Neuronen für den Fehler in der Ausgabe. Anhand dieses Fehlers des Neurons j kann das Netz dann die Anpassung der Verbindungsgewichte vornehmen, damit die verdeckten Neuronen den korrekten Aktivierungswert bekommen, der dann zusammen genommen den richtigen Ausgabewert ergibt.[114] Diese Anpassung der Gewichte erfolgt rückwärts von der Ausgabeschicht, über die verborgenen Schichten, bis zur Eingabeschicht.[115]

---

[108] Vgl. dazu *Ertel*, Grundkurs Künstliche Intelligenz, S. 290; *Lämmel/Cleve*, Künstliche Intelligenz, Wissensverarbeitung - Neuronale Netze, S. 205 ff.

[109] Vgl. ausführlich zum Backpropagation-Algorithmus, dessen genaue Berechnung für das Verständnis der Arbeit nicht notwendig ist: *Ertel*, Grundkurs Künstliche Intelligenz, S. 291 ff.; *Lämmel/Cleve*, Künstliche Intelligenz Wissensverarbeitung - Neuronale Netze, S. 211 ff.. .

[110] Siehe dazu *Ertel*, Grundkurs Künstliche Intelligenz, S. 324 ff.

[111] Vgl. dazu *Ertel*, Grundkurs Künstliche Intelligenz, S. 270 f.; *Flasiński*, Introduction to Artificial Intelligence, S. 166.

[112] *Ertel*, Grundkurs Künstliche Intelligenz, S. 291.

[113] *Lämmel/Cleve*, Künstliche Intelligenz, Wissensverarbeitung - Neuronale Netze, S. 213.

[114] Zum Ganzen *Bermúdez*, Cognitive science: an introduction to the science of the mind, S. 138.

[115] *Lämmel/Cleve*, Künstliche Intelligenz, Wissensverarbeitung - Neuronale Netze, S. 213; *Ertel*, Grundkurs Künstliche Intelligenz, S. 293.

Neben der Netzarchitektur beeinflussen auch die Wahl des Lernverfahrens und die Anzahl der Trainingsdurchläufe die Leistungsfähigkeit des Netzes.[116] Der passende Lernalgorithmus muss nicht selten durch aufwendige Experimente ermittelt werden.[117]

G. Beispiel: Buchstabenerkennung

Die vorstehenden theoretischen Grundlagen sollen anhand eines einfachen Beispiels verdeutlicht werden.[118] Ein typisches Einsatzgebiet für vorwärtsgerichtete mehrlagige Netze, die mit einem Backpropagation-Algorithmus trainiert wurden, ist die Erkennung und Klassifizierung von Buchstaben.[119]
Ein zu diesem Zweck entworfenes und trainiertes Netz ist letters.net.[120] Das Netz besteht aus drei Schichten, mit 35 Neuronen in der Eingabeschicht, 10 Neuronen in der verborgenen Schicht und 26 Neuronen in der Ausgabeschicht.[121] Dieser letzte Wert lässt schon auf die Logik hinter dem Entwurf der Netzarchitektur schließen: Für jeden Buchstaben im lateinischen Alphabet gibt es in der Ausgabeschicht ein Neuron, das entsprechend aktiviert wird.
Die Anzahl der Neuronen in der ersten Schicht wiederum folgt aus einem Rasterbild, das Ausgangspunkt für die optische Zeichenerkennung ist. Für letters.net bilden die 35 Neuronen ein solches Rasterbild mit 5x7 Bildpunkten nach (siehe **Abb. 4**), mit dem alle Großbuchstaben des lateinischen Alphabets dargestellt werden können.[122] Das so entworfene Netz wird dann mit einem

---

[116] Vgl. *Lämmel/Cleve*, Künstliche Intelligenz, Wissensverarbeitung - Neuronale Netze, S. 241.
[117] *Ehinger/Stiemerling*, CR 2018, 761 (763).
[118] Anspruchsvollere Beispiele zum Kreditscoring mit einem mehrschichtigen, forwärtsgerichteten Netz finden sich bei *West*, Computers & Operations Research 2000, 1131 (1134 ff.); *Lämmel/Cleve*, Künstliche Intelligenz, Wissensverarbeitung - Neuronale Netze, S. 236 ff.
[119] *Lämmel/Cleve*, Künstliche Intelligenz, Wissensverarbeitung - Neuronale Netze, S. 222 f.
[120] Das Netz war als Beispiel im Java Neural Network Simulator (JavaNNS) der Universität Tübingen enthalten, https://ra.cs.uni-tuebingen.de/SNNS/welcome.html. Der Simulator ist zwar mittlerweile überholt und wird nicht mehr betreut. Das Netz ist jedoch sehr übersichtlich und für die hier anvisierte möglichst einfache Veranschaulichung vollkommen ausreichend. Für die zeitgemäße Anwendung von KNN sei auf die Open-Source Plattform TensorFlow von Google verwiesen, https://www.tensorflow.org/?hl=de (Webseiten zuletzt abgerufen am 26.10.2023)
[121] Vgl. *Lämmel/Cleve*, Künstliche Intelligenz, Wissensverarbeitung - Neuronale Netze, S. 223.
[122] *Lämmel/Cleve*, Künstliche Intelligenz, Wissensverarbeitung - Neuronale Netze, S. 223.

Backpropagation-Algorithmus trainiert und dann auf seine Generalisierungsfähigkeit getestet.[123] Auf diese Weise wird sichergestellt, dass das Netz auch bei Pixel-Störungen noch die richtigen Buchstaben erkennen kann.[124]

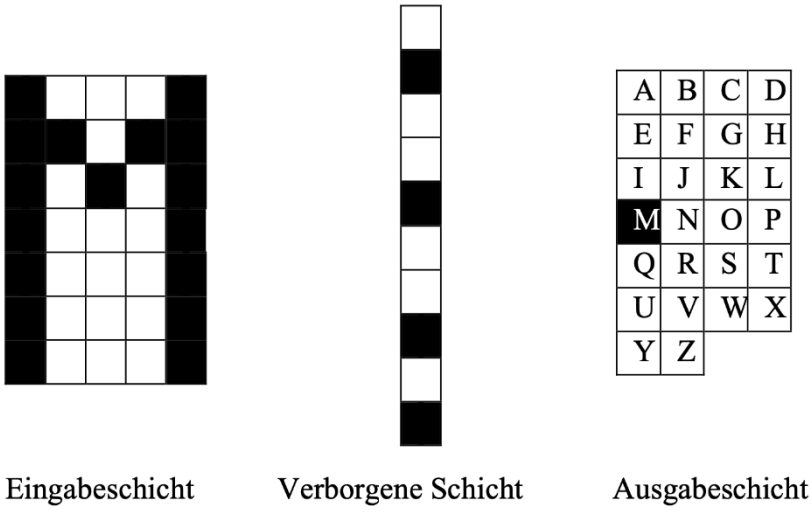

Abb. 4 *Das Netz letters.net. Abbildung nach Lämmel/Cleve.*[125]

H. Deep Learning

Wer sich mit der Frage von Transparenz im Bereich Künstliche Intelligenz beschäftigt, der begegnet schnell dem Begriff des *Deep Learning* („Tiefes Lernen"). Als *Deep Learning* werden Verfahren bezeichnet, bei denen sehr große – auch als „tief" bezeichnete – künstliche neuronale Netze mit großen Datensätzen trainiert werden.[126]

Erst durch Deep Learning konnten ML-Anwendungen in Bereichen erfolgreich entworfen werden, die lange Zeit als größte Herausforderung in der Entwicklung Künstlicher Intelligenz galten, wie etwa Spracherkennung und Erkennung

[123] Vgl. zu den Trainingsergebnissen mit unterschiedlichen Lernalgorithmen *Lämmel/Cleve*, Künstliche Intelligenz, Wissensverarbeitung - Neuronale Netze, S. 224 ff.

[124] Vgl. *Lämmel/Cleve*, Künstliche Intelligenz, Wissensverarbeitung - Neuronale Netze, S. 227.

[125] *Lämmel/Cleve*, Künstliche Intelligenz, Wissensverarbeitung - Neuronale Netze, S. 223.

[126] *Schaaf/Huber*, in: Bitkom, Blick in die Blackbox. Nachvollziehbarkeit von KI-Algorithmen in der Praxis, S. 62.

von handschriftlichen Dokumenten.[127] Möglich wurde dies, indem die Erstellung des Merkmalsvektors, also der Eingabedaten für ein künstliches neuronales Netz, nicht mehr manuell erfolgte, sondern bereits durch das Netz selbst (sog. *representation learning*).[128] Dafür wird üblicherweise einem überwacht lernenden Netz ein unüberwacht lernendes Netz vorgeschaltet, in dessen Schichten die Merkmalsextraktion stattfindet.[129]
Auf die Details zu Aufbau und Funktionalität von Netzen aus dem Bereich *Deep Learning* soll hier nicht weiter eingegangen werden.[130] Für die vorliegende Untersuchung ist vornehmlich die Frage der Skalierung von Interesse. Während Architektur und mithin auch Informationsverarbeitung in einem kleinen Netz noch nachvollziehbar gestaltet sein können, ist dies aufgrund der Größe und Komplexität beim *Deep Learning* nicht mehr der Fall.

I. Untersuchungsgegenstand: Trainiertes Netz

Künstliche Neuronale Netze können sowohl als untrainiertes als auch als trainiertes Modell vorliegen. Wurden die Parameter der Netzarchitektur wie Anzahl und Größe der Schichten, Art und Anzahl der Verbindungen zwischen den Schichten, sowie das Lernverfahren festgelegt, liegt zunächst einmal ein untrainiertes Modell vor. Diese Auswahlentscheidungen, die durch den Programmierer selbst gesetzt und nicht durch das Modell gelernt werden, werden als Hyperparameter bezeichnet.[131]
Das untrainierte Modell wird dann mit einem Set von Trainings- und Testdaten mit dem vorgesehenen Lernalgorithmus trainiert, Hyperparameter werden angepasst, Fehler gesucht und schließlich wird das Modell auf seine Generalisierungsfähigkeit getestet.[132] Das Ergebnis ist das „trainierte" Modell mit den der zu lösenden Aufgabe angepassten Gewichten.

---

[127] *Bermúdez*, Cognitive science: an introduction to the science of the mind, S. 318.
[128] Vgl. *Söbbing*, MMR 2021, 111 (112).
[129] Siehe dazu *Ertel*, Grundkurs Künstliche Intelligenz, S. 299 ff.
[130] Siehe dazu etwa *LeCun/Bengio/Hinton*, Nature 2015, 436 (436 ff.).
[131] *Gesellschaft für Informatik*, Technische und rechtliche Betrachtungen algorithmischer Entscheidungsverfahren, S. 54; *Drexl u. a.*, Technical Aspects of Artificial Intelligence: An Understanding from an Intellectual Property Law Perspective, Max Planck Institute for Innovation & Competition Research Paper No. 19-13, S. 6.
[132] *Russell/Norvig*, Artificial intelligence, S. 709.

Das trainierte Modell kann dann entweder während seines Einsatzes in diesem Zustand verbleiben (*offline*) oder es lernt durch die Daten, die es während seines Einsatzes erhält, weiter hinzu (*online*).[133]
Die vorliegende Arbeit legt den Fokus auf das trainierte (*offline*) Modell. Zwar müssen bereits bei der Programmierung der Neuronen, der Auswahl der Netzarchitektur und der Trainingskonfigurationen richtungsgebende Entscheidungen getroffen werden. Auch erfordert es Erfahrung und eine Vielzahl von Experimenten, um die besten Hyperparameter für das zu lösende Problem zu finden.[134] Sie haben maßgeblichen Einfluss darauf, wie genau das KNN in den Trainingsdaten komplexe Muster oder Anomalien erkennen kann.[135] Das untrainierte Netz hat also bereits einen nicht zu verachtenden wirtschaftlichen Wert. Dennoch ist etwa die Netzarchitektur von geringerem „strategischen Wert" als das trainierte Netz, weshalb sie auch häufig zugekauft beziehungsweise als open source offengelegt wird.[136]
Zudem wirken sich die beschriebenen Experimente zur Bestimmung der Hyperparameter auch unmittelbar auch auf das trainierte Modell aus, in das noch die Ergebnisse des Trainings und Testens mit einer oft umfangreichen Menge an Daten eingeflossen sind.[137]
Der Wert des trainierten Modells ergibt sich somit aus den Hyperparametern und den gelernten Parametern zugleich:

> „Aus wirtschaftlicher Sicht ist dieses Trainingsergebnis das monetarisierbare Ergebnis der Summe aller vorherigen Anstrengungen [...]."[138]

---

[133] Vgl. *Drexl u. a.*, Technical Aspects of Artificial Intelligence: An Understanding from an Intellectual Property Law Perspective, Max Planck Institute for Innovation & Competition Research Paper No. 19-13, S. 6; *Hartmann/Prinz*, in: Taeger, Rechtsfragen digitaler Transformationen - Gestaltung digitaler Veränderungsprozesse durch Recht, S. 772.

[134] *Ehinger/Stiemerling*, CR 2018, 761 (763 f.); *Russell/Norvig*, Artificial intelligence, S. 710.

[135] *Gesellschaft für Informatik*, Technische und rechtliche Betrachtungen algorithmischer Entscheidungsverfahren, S. 54.

[136] *Hartmann/Prinz*, in: Taeger, Rechtsfragen digitaler Transformationen - Gestaltung digitaler Veränderungsprozesse durch Recht, S. 776.

[137] Vgl. *Ehinger/Stiemerling*, CR 2018, 761 (764).

[138] *Ehinger/Stiemerling*, CR 2018, 761 (764).

J. Fazit

In diesem Kapitel wurden die grundlegenden Bestandteile eines Künstlichen Neuronalen Netzes, die Arten seines Trainings und seine Funktionsweise beschrieben.
Die einzelnen Neuronen bestehen aus mathematischen Funktionen und sind durch Gewichte verbunden. Die Anzahl der Neuronen und die Art der Verbindungen bestimmen die Netzarchitektur. Das so entstandene Netz wird mit Testdaten durch einen Lernalgorithmus trainiert und die Gewichte und gegebenenfalls der Schwellwert werden angepasst. Diese angepassten Gewichte machen somit das eigentliche Trainingsergebnis aus.
Nach der Lektüre dieses ersten Teils sollte auch deutlich geworden sein, dass KNN nicht die Unbeherrschbarkeit aufweisen, die Ihnen häufig zugeschrieben wird. Auf identische Eingaben liefert ein nicht mehr lernfähiges, also „eingefrorenes" Netz grundsätzlich identische Ausgaben.[139]
Auf der hier geschaffenen Grundlage kann im weiteren Verlauf untersucht und bewertet werden, welche Information die einzelnen Bestandteile des Netzes enthalten, wie sie dargestellt werden können und inwiefern ihre Auswahl, Kombination und Implementierung relevant ist für die immaterialgüterrechtliche Schutzfähigkeit eines trainierten Künstlichen Neuronalen Netzes.

[139] *Zech*, Weizenbaum Series 2020, 1 (47).

Teil 2

Information, Blackbox und Transparenz

Kapitel 2

Informationsbegriffe

Anknüpfungspunkt für den Schutz als Geschäftsgeheimnis ist „Information". Was genau unter diesem allgegenwärtigen Begriff im Rahmen des Geheimnisschutzrechts verstanden wird, wird in diesem Kapitel untersucht. Doch die rechtliche Definition soll hier nicht für sich stehen – vielmehr werden weitere Definitionen und Kategorien von Information untersucht, welche die spätere Analyse des Geschäftsgeheimnisses an einem Künstlichen Neuronalen Netz unterstützen und erleichtern werden.

A. Geschäftsgeheimnisgesetz

Der Begriff der Information bildet die Grundlage der Definition eines Geschäftsgeheimnisses gemäß § 2 Nr. 1 des Gesetzes zum Schutz von Geschäftsgeheimnissen (GeschGehG)[140], an den alle weiteren Tatbestandsmerkmale der Buchstaben a - c anknüpfen. Wer sich aus dem GeschGehG jedoch eine Definition von „Information" erhofft, wird enttäuscht. Es handelt sich um ein sehr weites Tatbestandsmerkmal, das erst durch die weiteren Tatbestandsmerkmale des Geschäftsgeheimnisses eingegrenzt wird.
Im Wege der Auslegung kann jedoch auf die Geschäftsgeheimnis-Richtlinie (Geschäftsgeheimnis-RL)[141] zurückgegriffen werden, die im GeschGehG in deutsches Recht umgesetzt wird. Nach Erwägungsgrund 2 der Geschäftsgeheimnis-RL umfassen Geschäftsgeheimnisse ein „breites Spektrum von Infor-

[140] Ein Auszug aus dem Geschäftsgeheimnisgesetz mit den hier analysierten Normen findet sich im Anhang.

[141] Richtlinie (EU) 2016/943 vom 8. Juni 2016 über den Schutz vertraulichen Know-hows und vertraulicher Geschäftsinformationen (Geschäftsgeheimnisse) vor rechtswidrigem Erwerb, sowie rechtswidriger Nutzung und Offenlegung.

mationen, das über das technologische Wissen hinausgeht und auch Geschäftsdaten wie Informationen über Kunden und Lieferanten, Businesspläne sowie Marktforschung und -strategien einschließt." Abstrakter gesprochen kann Information im Sinne des GeschGehG „Angaben, Daten, Kommunikationsakte, Umstände oder sonstiges Wissen"[142] umfassen.
Der Hinweis auf „sonstiges Wissen" zeigt, dass Information im Sinne des GeschGehG auf semantischer Ebene abgegrenzt wird.[143] Das bedeutet jedoch nicht, dass nur Information, deren Bedeutung für den menschlichen Betrachter unmittelbar verfügbar ist, also faktische semantische Information, ein Geschäftsgeheimnis im Sinne des Gesetzes sein kann.[144] Eine derart enge Auslegung würde den Geheimnisschutz unangemessen eingrenzen und wäre auch nicht handhabbar, da der Übergang von potenzieller zu faktischer Information subjektiv und fließend ist. Bisweilen wird vertreten, dass auch Information ohne semantische Ebene, mithin reine syntaktische Information, als Geschäftsgeheimnis geschützt sein kann beziehungsweise sollte.[145] Ausschlaggebend für diese Überlegung ist der große wirtschaftliche Wert von Daten, dem auch ein moderner Geheimnisschutz Rechnung tragen sollte.[146] Gerade die sogenannten Rohdaten sollen nicht aus dem Schutz des Geschäftsgeheimnisgesetzes herausfallen. Nach hiesigem Verständnis ergibt sich der Schutz von Rohdaten als Geschäftsgeheimnis jedoch nicht daraus, dass auch rein syntaktische Information durch das GeschGehG geschützt wäre. Vielmehr haben auch Rohdaten eine semantische Ebene, denn auch sie sind immer durch ihre Erhebung kontextualisiert.[147] Als potenzielle semantische Information können sie mithin auch Schutzgegenstand des GeschGehG sein.

---

[142] *Alexander*, in: Köhler/Bornkamm/Feddersen, GeschGehG, § 2 Rn. 25.

[143] Siehe noch zum alten Recht *Zech*, Information als Schutzgegenstand, S. 230 f.; *Zech*, JIPLP 2016, 460 (465); zum neuen Recht *Zech*, GRUR 2015, 1151 (1156).

[144] Siehe zur Unterscheidung von potenzieller und faktischer Information Abschnitt C dieses Kapitels.

[145] *Hauck*, NJW 2016 2218 (2221); *Sagstetter*, in: Maute/Mackenrodt, Recht als Infrastruktur für Innovation, S. 292; *Hoppe*, in: Hoppe/Oldekop, Geschäftsgeheimnisse. Schutz von Know-how und Geschäftsinformationen. Praktikerhandbuch mit Mustern, S. 23; siehe zur Unterscheidung von semantischer und syntaktischer Information Abschnitt D dieses Kapitels.

[146] Siehe zum Schutz von Daten in der Geschäftsgeheimnis-Richtlinie *Zech*, JIPLP 2016, 460, S. 465; *Alexander*, WRP 2017, 1034 (1038); *Sagstetter*, in: Maute/Mackenrodt, Recht als Infrastruktur für Innovation, S. 292 ff.

[147] Zur Kontextualisierung von Rohdaten in Datensets siehe *Sagstetter*, in: Maute/Mackenrodt, Recht als Infrastruktur für Innovation, S. 292 f.

Die Verwendung des Begriffs „Information" im Singular durch das Geschäftsgeheimnisgesetz ist darüber hinaus ungenau: bei der Information als Grundlage eines Geschäftsgeheimnisses wird es sich meist um eine Menge von Einzelinformationen handeln.[148] Das verdeutlicht bereits der Wortlaut des § 2 Nr. 1a GeschGehG, nach dem die geschützte Information „weder insgesamt noch in der genauen Anordnung und Zusammensetzung ihrer Bestandteile" allgemein bekannt oder ohne Weiteres zugänglich sein darf.[149] Für den Schutz einer Kombination von Einzelinformationen als Geschäftsgeheimnis ist es auch nicht erforderlich, dass die Einzelinformationen für sich allein schutzfähig sind.[150]
Anders als das Urheberrecht nimmt das Geschäftsgeheimnisrecht keine Einschränkung im Hinblick auf den Erzeuger des Schutzgegenstands vor. Information im Sinne des GeschGehG muss keine „persönliche geistige Schöpfung" im Sinne des § 2 Abs. 2 UrhG sein, sie kann auch von einem Computer erzeugt worden sein.[151] Ebenso wenig werden besondere Anforderungen an die Qualität der Information gestellt, sie muss weder innovativ sein noch einen (aktuellen) Nutzungswert haben.[152] Die Geschäftsgeheimnis-RL schließt zwar in Erwägungsgrund 14 „belanglose Information" von ihrem Schutzbereich aus. Allerdings ist diese Einschränkung bereits im weiteren Tatbestandsmerkmal des wirtschaftlichen Wertes enthalten und dürfte daher nicht schon den weiten Informationsbegriff des GeschGehG einschränken.[153]
Auch auf die Art der Verkörperung der Information kommt es nicht an, sie kann sowohl physisch als auch rein virtuell vorliegen.[154] Nur muss sie in irgendeiner Art und Weise verkörpert sein, bloße Gedanken können nicht Gegenstand von Geheimhaltungsmaßnahmen sein und fallen daher nicht unter Informationen im Sinne des GeschGehG.[155]

---

[148] *Harte-Bavendamm*, in: Harte-Bavendamm/Ohly/Kalbfus, GeschGehG, § 2 Rn. 14.

[149] Zu diesen und weiteren Merkmalen des Geschäftsgeheimnisses unten, Kapitel 4.

[150] Siehe nur *Harte-Bavendamm*, in: Harte-Bavendamm/Ohly/Kalbfus, GeschGehG, § 2 Rn. 14.

[151] *Alexander*, in: Köhler/Bornkamm/Feddersen, GeschGehG, § 2 Rn. 26; *Hessel/Lesser*, MMR 2020, 647 (649).

[152] *Harte-Bavendamm*, in: Harte-Bavendamm/Ohly/Kalbfus, GeschGehG, § 2 Rn. 20; *Alexander*, in: Köhler/Bornkamm/Feddersen, GeschGehG, § 2 Rn. 27 f.

[153] so auch *Alexander*, in: Köhler/Bornkamm/Feddersen, GeschGehG, § 2 Rn. 29; *Hoppe*, in: Hoppe/Oldekop, Geschäftsgeheimnisse. Schutz von Know-how und Geschäftsinformationen. Praktikerhandbuch mit Mustern, S. 23.

[154] *Alexander*, in: Köhler/Bornkamm/Feddersen, GeschGehG, § 2 Rn. 25.

[155] *Alexander*, in: Köhler/Bornkamm/Feddersen, GeschGehG, § 2 Rn. 26.

Um als Geschäftsgeheimnis geschützt zu sein, muss die Information noch die weiteren Voraussetzungen des § 2 Nr. 1 a-c GeschGehG erfüllen. Sie darf weder insgesamt noch in der genauen Anordnung und Zusammensetzung ihrer Bestandteile den Personen in den Kreisen, die üblicherweise mit dieser Art von Informationen umgehen, allgemein bekannt oder ohne Weiteres zugänglich sein (lit. a). Aus dieser Geheimhaltung muss sich darüber hinaus ein wirtschaftlicher Wert ergeben (lit. a am Ende). Die Information muss außerdem Gegenstand von den Umständen nach angemessenen Geheimhaltungsmaßnahmen durch ihren rechtmäßigen Inhaber sein (lit. b) und es muss ein berechtigtes Interesse an ihrer Geheimhaltung bestehen (lit. d). Diese weiteren Tatbestandsmerkmale des Geschäftsgeheimnisses werden im Rahmen der Prüfung der Schutzfähigkeit von KNN im dritten Teil eingehend erläutert.

B. Information als Gegenteil von Unbestimmtheit

Nach dem wohl grundlegendsten Informationsbegriff ist Information das „Gegenteil von Unbestimmtheit“[156] beziehungsweise, dynamisch gedacht, die „Verminderung der Unbestimmtheit des Zustandes eines Systems.“[157] Dem liegt die Überlegung zu Grunde, dass Information immer eine „gewisse Unbestimmtheit“ über die Auswahl aus einer Quelle von Information voraussetzt.[158]
Der Informationsgehalt ist dann quantifizierbar, wenn er als äquivalent zum Maß der Unbestimmtheit ausgedrückt wird:

> „Je mehr Zustände möglich sind, ohne dass eine Aussage darüber getroffen werden kann, welchen Zustand das System einnimmt, desto unbestimmter ist das System, desto geringer die Information über das System. Umgekehrt lässt sich sagen, dass ein System, das einen bestimmten Zustand aus mehreren möglichen eingenommen hat, ein bestimmtes Maß an Information enthält.“[159]

Wie viel Information ein System enthält, ist dann eine Frage der Wahrscheinlichkeit: je wahrscheinlicher der eingenommene Zustand zu erwarten war, desto

---

[156] *Zech*, Information als Schutzgegenstand, S. 14, m.w.N.
[157] *Ebeling/Freund/Schweitzer*, Komplexe Strukturen, S. 40.
[158] *Schönfeld/Klimant/Piotraschke*, Informations- und Kodierungstheorie, S. 11.
[159] *Zech*, Information als Schutzgegenstand, S. 19.

geringer ist der Informationsgehalt, je unwahrscheinlicher der eingenommene Zustand war, desto größer ist der Informationsgehalt. Das Maß der Information ist also der „Kehrwert der Wahrscheinlichkeit", dass ein bestimmter Zustand eingenommen wird.[160]
Das Maß der Unbestimmtheit eines Systems wird in der Informationstheorie auch als Entropie[161] bezeichnet. Der Begriff stammt aus der Thermodynamik und bezeichnet dort das Maß der Unordnung eines Systems. Aufgrund des Zusammenhangs von Unbestimmtheit und Information wird er in der Informationstheorie zur Beschreibung des mittleren Informationsgehalts der Zeichen eines Systems und daher des Informationsgehalts dieses Systems verwendet.[162] Die Entropie beschreibt dann die Wahrscheinlichkeit, dass sich ein bestimmtes System realisiert. Diese Wahrscheinlichkeit hängt von der Zahl der Möglichkeiten ab, durch die das System realisiert werden kann: sind diese Möglichkeiten sehr zahlreich, ist die Entropie des Systems groß und der Informationsgehalt gering.[163]

C. Potenzielle und faktische Information

Eine weitere zur Beschreibung von Information getroffene Unterscheidung kann bei der vorliegenden Analyse helfen: die Differenzierung zwischen potenzieller und faktischer Information.[164] Dabei wird der Beobachter in den Blick genommen: potenzielle Information stellt dann die „Gesamtheit aller Aussagen" die zur Beschreibung eines Systems erforderlich sind dar, während faktische Information die tatsächliche durch den Beobachter getroffenen Aussagen darstellt. Dadurch kann „die beobachtbare Struktur eines Systems mit der in der Struktur enthaltenen potenziellen Information gleich[ge]setz[t]"[165] werden.

---

[160] Vgl. zu diesem Absatz *Zech*, Information als Schutzgegenstand, S. 18 f.; *Schönfeld/Klimant/Piotraschke*, Informations- und Kodierungstheorie, S. 12.

[161] Vgl. zum Begriff *Ebeling/Freund/Schweitzer*, Komplexe Strukturen, S. 29 ff.; *Zech*, Information als Schutzgegenstand, S. 20 f.

[162] *Schönfeld/Klimant/Piotraschke*, Informations- und Kodierungstheorie, S. 16; *Zech*, Information als Schutzgegenstand, S. 20.

[163] *Ebeling/Freund/Schweitzer*, Komplexe Strukturen, S. 40; *Zech*, Information als Schutzgegenstand, S. 20.

[164] Vgl. zu diesem Absatz *Zech*, Information als Schutzgegenstand, S. 15 f.

[165] *Zech*, Information als Schutzgegenstand, S. 16.

Aus der Kombination der Konzepte von Information als „beseitigte Unbestimmtheit“[166] und als potenzielle und faktische Information lässt sich folgern und weiter differenzieren: ein System mit sehr vielen möglichen Zuständen hat ein hohes Maß potenzieller Information. Lässt sich diese Information jedoch aufgrund von Größe und/oder Komplexität des Systems nur schwer umschreiben, ist die faktische Information dennoch gering.[167]
Auch die Komplexität eines Systems, also die „Anzahl von Elementen und Relationen“[168], gibt Aufschluss über seinen Informationsgehalt.[169] Je komplexer und asymmetrischer ein System, desto höher ist sein potenzieller Informationsgehalt. Eine Quantifizierung der Komplexität eines Systems ist durch die sogenannte algorithmische Komplexität möglich, also die „Länge des kürzesten Programms [...], das dieses System reproduziert.“[170] Bei reiner Betrachtung der faktischen Information ist die algorithmische Komplexität dann die Komplexität des Systems selbst.[171]

D. Semantische, syntaktische und strukturelle Information

Neben dieser grundsätzlichen Charakterisierung von Information als „Gegenteil von Unbestimmtheit“, die zur Beschreibung des Informationsgehalts eines Systems und zur Differenzierung zwischen potenzieller und faktischer Information führt, können auch verschiedene Arten von Information unterschieden werden. Für diese Untersuchung besonders hilfreich ist die Unterscheidung von semantischer, syntaktischer und struktureller Information, wobei auch von der semantischen, syntaktischen oder strukturellen Ebene von Information gesprochen werden kann. Die erste bezieht sich auf die Bedeutungsebene, die zweite auf die Zeichenebene, die dritte auf die körperliche Ebene.[172] Für die Betrachtung der Information eines KNN sind insbesondere die semantische und

[166] *Schönfeld/Klimant/Piotraschke*, Informations- und Kodierungstheorie, S. 12.
[167] *Zech*, Information als Schutzgegenstand, S. 19 f.
[168] *Zech*, Information als Schutzgegenstand, S. 21.
[169] zu diesem Absatz *Zech*, Information als Schutzgegenstand, S. 21.
[170] *Ebeling/Freund/Schweitzer*, Komplexe Strukturen, S. 25.
[171] *Zech*, Information als Schutzgegenstand, S. 21.
[172] Siehe dazu *Zech*, Information als Schutzgegenstand, S. 3 m.w.N., 37 ff.

die syntaktische Ebene von Interesse. Das Konzept lässt sich am besten durch ein Beispiel verdeutlichen: ein mit Buchstaben beschriebenes Blatt Papier ist aufgrund der Verkörperung strukturelle Information und beinhaltet in jedem Fall syntaktische Information, denn diese findet sich auf der Zeichenebene. Diese syntaktische Information muss jedoch nicht zwingend auch semantische Information enthalten. Handelt es sich um eine wahllose Aneinanderreihung von Buchstaben, die keiner Regel (also keinem Code) folgt, so handelt es sich um rein syntaktische Information. Haben die Buchstaben jedoch in einer Sprache oder nach einem anderen Code für einen Leser eine Bedeutung, bergen sie auch semantische Information.

E. Implizite und explizite Information

Die Entwicklung Künstlicher Neuronaler Netze beruht auf der Simulation von Prozessen im menschlichen Gehirn. Maßgeblich sowohl für Biologie als auch für Informatik ist dabei ein Verständnis von der Verarbeitung und Darstellung von Information. Dies führte zur Herausbildung der Kognitionswissenschaften, die sich mit der Informationsverarbeitung in beiden Bereichen befasst.[173]
Die Kognitionswissenschaft unterscheidet implizites und explizites Wissen. Während letzteres auf einer bewussten Aneignung von Daten beruht, wird implizites Wissen im Wesentlichen unbewusst durch sensorisches und motorisches Lernen erworben.[174] Explizites Wissen kann umschrieben werden als „Wissen, das nicht nur irgendwie vorhanden ist, sondern mittels Sprache oder Bildern (allgemein Zeichen) so ausgedrückt werden kann, dass eine Kommunikation möglich wird."[175] Daran anknüpfend werden auch informationsverarbeitende Systeme unterschieden, die Information symbolisch repräsentieren und solche, die sie implizit repräsentieren.[176]
Frühe Vertreter der KI-Forschung gingen noch davon aus, dass jedes intelligente System auf der symbolischen Repräsentation von Wissen beruhe und dass das

[173] *Zech*, Weizenbaum Series 2020, 1 (8).
[174] *Mainzer*, Künstliche Intelligenz – Wann übernehmen die Maschinen?, S. 102.
[175] *Lämmel/Cleve*, Künstliche Intelligenz, Wissensverarbeitung - Neuronale Netze, S. 25.
[176] *Mainzer*, Künstliche Intelligenz – Wann übernehmen die Maschinen?, S. 102; *Zech*, Weizenbaum Series 2020, 1 (7); siehe für einen ausführlichen Vergleich der Wissensrepräsentation in GOFAI und KNN *Haykin*, Neural Networks: A comprehensive Foundation, S. 56 ff.

Wissen in einem intelligenten System daher explizit sei. Bereits in den 1950er Jahren legten zwei Pioniere der KI-Forschung, Allen Newell und Herbert Simon, mit der Entwicklung von auf symbolischer Repräsentation beruhenden Computerprogrammen den Grundstein dieser Annahme, die sie später in ihrer *Physical Symbol System Hypothesis* festschrieben.[177] Ihre Hypothese lautete:

> "A physical symbol system has the necessary and sufficient means for general intelligent action."[178]

Für Newell und Simon war ein System von Symbolen notwendige und hinreichende Bedingung für Intelligenz. Nach ihrer Hypothese könnten mithin alle kognitiven Fähigkeiten durch symbolische Repräsentation beschrieben werden und beruhten ihrerseits notwendigerweise auf symbolischer Repräsentation. Unter einem physikalischen Symbolsystem verstanden sie folgendes:

> "A physical symbol system consists of a set of entities, called symbols, which arc physical patterns that can occur as components of another type of entity called an expression (or symbol structure). Thus, a symbol structure is composed of a number of instances (or tokens) of symbols related in some physical way (such as one token being next to another). At any instant of time the system will contain a collection of these symbol structures. Besides these structures, the system also contains a collection of processes that operate on expressions to produce other expressions: processes of creation, modification, reproduction and destruction. A physical symbol system is a machine that produces through time an evolving collection of symbol structures. Such a system exists in a world of objects wider than just these symbolic expressions themselves."[179]

Ein physikalisches Symbolsystem hat demnach folgende Eigenschaften: es besteht aus Symbolen, das heißt physikalischen Mustern, und Symbolstrukturen. Außerdem verfügt das System über Prozesse, um Symbolstrukturen zu erschaffen, zu verändern, zu vervielfältigen oder zu löschen. Dies entspricht einem klassischen Computerprogramm, in dem jedem Zeichen eine Bedeutung zugewie-

---

[177] *Bermúdez*, Cognitive science: an introduction to the science of the mind, S. 100 ff.; *Mainzer*, Künstliche Intelligenz – Wann übernehmen die Maschinen?, S. 143.

[178] *Newell/Simon*, Communications of the ACM 1975, 113 (116).

[179] *Newell/Simon*, Communications of the ACM 1975, 113 (116).

sen ist und mit den Zeichen logisch operiert wird, ebenso wie bei der klassischen Künstlichen Intelligenz (*Good Old-Fashioned AI*, GOFAI).[180]
Doch die These, dass Denken und Intelligenz letztendlich in der Veränderung von Symbolen anhand bestimmter Regeln beruhen, wird auch als verkürzt kritisiert.[181] Künstliche Neuronale Netze zeigen, dass Wissen nicht klar abgrenzbar durch bestimmte Symbole in einem System repräsentiert und auffindbar, mithin explizit, sein muss.[182] Implizites Wissen benötigt keine „regelbasierte Repräsentation".[183] Zwar gibt es KNN, in denen jedes Neuron eine bestimmte Eigenschaft der Eingabedaten repräsentiert. Solche KNN werden als lokale Netze (*localist connectionist networks*) bezeichnet.[184] Der für Wissenschaft und Anwendung wesentliche reizvolleren KNN sind jedoch sog. *distributed connectionist networks*, in denen Wissen auf verschiedene Orte verteilt, in den gewichteten Verbindungen zwischen einzelnen Neuronen gespeichert wird. Das enthaltene Wissen beziehungsweise die enthaltene Information kann damit nicht an spezifischen Punkten lokalisiert werden.[185] Die im ersten Teil der Arbeit beschriebenen KNN sind solche *distributed networks*:

> „Der konnektionistische Ansatz betont deshalb, dass Bedeutung nicht von Symbolen getragen wird, sondern sich in der Wechselwirkung zwischen verschiedenen kommunizierenden Einheiten eines komplexen Netzwerks ergibt. Diese Herausbildung bzw. Emergenz von Bedeutungen und Handlungsmustern wird durch die sich selbst organisierende Dynamik von neuronalen Netzwerken [...] möglich."[186]

---

[180] *Zech*, Weizenbaum Series 2020, 1 (13 f.); siehe zur „Symbolischen KI" auch *Flasiński*, Introduction to Artificial Intelligence, S. 15 ff.

[181] Als Gegenbeweis erdachte der Philosoph John Searle etwa das sog. Chinese room argument, siehe dazu *Bermúdez*, Cognitive science: an introduction to the science of the mind, S. 117 f.; siehe auch *Mainzer*, Künstliche Intelligenz – Wann übernehmen die Maschinen?, S. 143; *Kaplan*, Künstliche Intelligenz, S. 37.

[182] *Bermúdez*, Cognitive science: an introduction to the science of the mind, S. 141.

[183] *Mainzer*, Künstliche Intelligenz – Wann übernehmen die Maschinen?, S. 207.

[184] *Bermúdez*, Cognitive science: an introduction to the science of the mind, S. 141; *Flasiński*, Introduction to Artificial Intelligence, S. 24.

[185] *Bermúdez*, Cognitive science: an introduction to the science of the mind, S. 141; *Flasiński*, Introduction to Artificial Intelligence, S. 24.

[186] *Mainzer*, Künstliche Intelligenz – Wann übernehmen die Maschinen?, S. 144.

Auch wenn die Kognitionswissenschaft das menschliche Gehirn und digitale Technologie verknüpft, muss eine Unterscheidung ausdrücklich hervorgehoben werden, um Missverständnissen vorzubeugen. Implizites Wissen ist Gegenstand der Psychologie und Kognitionsforschung und meint dasjenige Wissen, das durch motorische oder sensorische Erfahrung gebildet wird und das nicht symbolisch repräsentiert und auf diese Weise angeeignet werden kann. Da Information, um maschinenlesbar zu sein, immer repräsentiert werden können muss, kann ein Computer kein implizites Wissen haben. Die Information kann jedoch implizit, das heißt ohne konkrete symbolische Repräsentation in einem Computerprogramm enthalten sein.[187] Dann wird teilweise vereinfacht auch von „implizitem Wissen" gesprochen.[188] Genauer ist es, im Bereich des Maschinellen Lernens von „Information" zu sprechen:

> „Information kann durch komplexe Systeme erzeugt werden, Wissen nur durch den menschlichen Geist."[189]

Um diese implizit repräsentierte Information explizit zu machen, muss sie mithilfe von Zeichen symbolisch repräsentiert werden. Ergebnis können ganz unterschiedliche Arten von expliziter Information sein, etwa Text, Daten oder Bilder.[190] Diese Repräsentation muss anhand einer bestimmten Regel erfolgen, die in der Semiotik als Code bezeichnet wird.[191] Im dritten Teil dieser Arbeit werden unterschiedliche solcher Codes zur Darstellung der impliziten Information eines KNN vorgestellt und untersucht.

---

[187] Dass Information in der GOFAI symbolisch repräsentiert und in neueren Formen, v.a. KNN, implizit repräsentiert ist, ist eine gängige Unterscheidung in der Kognitionswissenschaft, vgl. nur *Flasiński*, Introduction to Artificial Intelligence, S. 15 ff., 23 ff.; *Mainzer*, Künstliche Intelligenz – Wann übernehmen die Maschinen?, S. 102 ff.; *Bermúdez*, Cognitive science: an introduction to the science of the mind, S. 99 ff., 141 ff., der jedoch nicht den Begriff „implizit" verwendet; weniger differenziert dagegen Konertz/Schönhof, nach denen KI kein implizites Wissen haben kann: *Konertz/Schönhof*, Das technische Phänomen „Künstliche Intelligenz" im allgemeinen Zivilrecht, S. 68 und 134.

[188] so etwa *Mainzer*, Künstliche Intelligenz – Wann übernehmen die Maschinen?, S. 102.

[189] *Zech*, Information als Schutzgegenstand, S. 33.

[190] *Zech*, Information als Schutzgegenstand, S. 39.

[191] *Zech*, Information als Schutzgegenstand, S. 24.

F. Fazit

Der Informationsbegriff des Geschäftsgeheimnisgesetzes ist ausgesprochen weit. Er sagt jedoch nichts über das Wesen der Information aus, vielmehr wird das Verständnis, was genau „Information" ist, schlicht vorausgesetzt. Der Blick in andere Disziplinen führt da schon etwas weiter: Information kann als Gegenteil von Unbestimmtheit beschrieben werden. Daran anknüpfend erschließt sich die Unterscheidung von potentieller Information, die grundsätzlich in einem System vorhanden ist, und faktischer Information, die tatsächlich durch einen Beobachter herausgefiltert wird. Für die juristische Analyse besonders hilfreich ist die Unterscheidung von semantischer, syntaktischer und struktureller Information, da anhand ihrer die immaterialgüterrechtlichen Schutzgegenstände identifiziert und abgegrenzt werden können. Abschließend hilft der Blick in die Kognitionswissenschaft und die Informatik, um mit der impliziten und der expliziten Information eine unmittelbar mit dem Untersuchungsgegenstand Künstliches Neuronales Netz verbundene Brücke zu potentieller und faktischer Information zu schlagen, die eine große Nähe aufweisen.

Kapitel 3

Blackbox und Transparenz

Künstliche Intelligenz wird in so gut wie allen Bereichen der Privatwirtschaft eingesetzt und hat auch in der öffentlichen Verwaltung und der Justiz Einzug erhalten.[192] Sie gestaltet die digitale Gesellschaft und damit auch vollumfänglich unseren Alltag.[193] Diese rasante Entwicklung wird flankiert von einem Unbehagen gegenüber dem technischen Fortschritt und der „Entscheidungsmacht" von Computern, das sich in Rufen nach Regulierung und meist auch nach mehr Transparenz ausdrückt.[194]

Das gesellschaftliche Unbehagen betrifft dabei vornehmlich die Entscheidungsfindung durch Künstliche Intelligenz und deren Intransparenz.[195] Denn neben dem großen Nutzen, die der Einsatz künstlicher Intelligenz mit sich bringen kann, bestehen auch vielfältige Risiken für den Einzelnen und die Gesellschaft.[196] Häu-

---

[192] In Deutschland steckt der Einsatz von KI im öffentlichen Sektor zwar noch in den Anfängen, die Europäische Kommission will ihn aber explizit fördern, *Europäische Kommission*, Weißbuch zur Künstlichen Intelligenz, S. 9; ein Blick in die USA bietet plausible Zukunftsszenarien, vgl. *Martini*, Blackbox Algorithmus – Grundfragen einer Regulierung Künstlicher Intelligenz, S. 86 ff.

[193] *Vieth/Wagner*, Teilhabe, ausgerechnet, S. 8.

[194] Siehe statt vieler *Martini*, Blackbox Algorithmus – Grundfragen einer Regulierung Künstlicher Intelligenz; *Baase*, A gift of fire; *Wischmeyer*, AöR 2018, 1; *Deutscher Bundestag*, Bericht der Enquete-Kommission Künstliche Intelligenz – Gesellschaftliche Verantwortung und wirtschaftliche, soziale und ökologische Potenziale, BT-Drs. 19/23700; *Bundesregierung*, Strategie Künstliche Intelligenz der Bundesregierung; *Europäische Kommission*, Weißbuch zur Künstlichen Intelligenz.

[195] „Das am meisten in der Diskussion genannte Problem mit algorithmischer Entscheidungsfindung ist der Mangel an Transparenz, die Angst vor der Blackbox." *Gesellschaft für Informatik*, Technische und rechtliche Betrachtungen algorithmischer Entscheidungsverfahren, S. 112.

[196] Sehr übersichtlich dargestellt sind Gefährdungslagen durch einige lernfähige Softwareanwendungen bei *Martini*, Blackbox Algorithmus – Grundfragen einer Regulierung Künstlicher Intelligenz, S. 110 ff.

figes Schlagwort ist dabei das „Vertrauen“ in KI. Um dieses herzustellen, befürwortet der Großteil der Interessenträger im Bereich KI eine entsprechende Regulierung.[197]
Die Regulierungsdebatte um Künstliche Intelligenz wird in der Rechtswissenschaft zwar schon seit einiger Zeit geführt, sie findet jedoch erst in den letzten Jahren ihren Niederschlag in der deutschen und der EU-Gesetzgebung. Erste Transparenzpflichten sind bereits in Kraft, andere befinden sich noch im Entwurfsstadium.
Da diese erste Phase der Regulierung noch lange nicht abgeschlossen ist, ist eine genaue Betrachtung der vorhandenen und im Entstehen befindlichen Transparenzpflichten müßig. Sie ist jedoch für die vorliegende Fragestellung auch nicht zwingend erforderlich, denn für den Geheimnisschutz kommt es auf wenige Parameter an, die über den Verlust von Geschäftsgeheimnissen entscheiden und die abstrakt, unabhängig von einer konkreten Regelung, analysiert werden können. Bei diesen Parametern handelt es sich um die schlichte Frage „was“, „an wen“ und „unter welchen Bedingungen“ offengelegt werden muss.
Bevor auf verschiedene Darstellungsformen der Informationen eines KNN und mithin die Möglichkeit der Erfüllung von Transparenzpflichten eingegangen wird, sollen zunächst der Blackbox-Charakter[198] des Netzes sowie die Transparenzdebatte dargestellt und anschließend einige Regulierungsbestrebungen umrissen werden.

A. Blackbox und Komplexität

Der Begriff „Blackbox“ wurde ursprünglich in der Regeltechnik verwendet und beschreibt ein System, von dem nur die Eingangs- und die Ausgangssignale bekannt sind.[199] Um ein solches System handelt es sich auf den ersten Blick auch bei vielen ML-Anwendungen. Dementsprechend verbreitet ist der Begriff Blackbox in der Debatte um Algorithmen beziehungsweise Künstliche Intelligenz. Dabei wird das Bild der Blackbox weniger in der Informatik selbst, sondern vielmehr in

[197] *Europäische Kommission*, Entwurf der KI-Verordnung, COM(2021) 206 final, S. 1, 9.

[198] Der Begriff wird so gebraucht von Martini: *Martini*, Blackbox Algorithmus – Grundfragen einer Regulierung Künstlicher Intelligenz, S. 136, 355; andere sprechen von „blackboxness“, *Lipton*, ICML WHI 2016, 96 (98).

[199] Siehe zum Begriff *Geitz/Vater/Zimmer-Merkle*, Black Boxes - Versiegelungskontexte und Öffnungsversuche: interdisziplinäre Perspektiven, S. 3 ff.

der rechts- und sozialwissenschaftlichen Diskussion gebraucht.[200] Allenfalls das Feld der XAI bildet naturgemäß eine Ausnahme.[201]

Neben dem „Blackbox-Charakter" von Algorithmen oder ML-Modellen wird auch von opaken Systemen beziehungsweise von Opazität gesprochen. Dieser Begriff wird auch vermehrt in der XAI-Forschung gebraucht.

Burrell unterscheidet drei Arten von algorithmischer Opazität: Opazität aufgrund von Geheimhaltung, Opazität mangels Programmierkenntnissen und Opazität aufgrund der Diskrepanz zwischen algorithmischer Komplexität und menschlicher Auffassungsgabe.[202] Sie verwendet den Begriff „opak" dabei sowohl hinsichtlich der Funktionsweise eines Systems als auch hinsichtlich der Begründung einer konkreten Entscheidung.[203] Der dritten Form von Opazität liegt die bereits beschriebene Unterscheidung zwischen potenzieller und faktischer Information zugrunde.

Denn anknüpfend an informationstheoretische Grundsätze kann die Architektur eines KNN, also die Anzahl der Neuronen, der Schichten und die Art und Anzahl der Verbindungen, auch als komplexe Struktur bezeichnet werden. Unter Struktur wird in der Informationstheorie die „Art der Zusammensetzung eines Systems aus Elementen und die Menge der Relationen bzw. Operationen, welche die Elemente miteinander verknüpfen", verstanden.[204] Eine Struktur ist komplex, „wenn sie durch Ordnungsrelationen bzw. Korrelationen auf vielen Skalen charakterisiert" wird.[205] Bei linearen Modellen kann Komplexität gemessen werden an der Anzahl der Gewichte, die nicht Null sind, bei Entscheidungsbäumen an ihrer Tiefe und bei Regeln an der Länge ihrer Bedingung, das heißt der Anzahl von Merkmalen und dazugehörigen Werten.[206] Die Tiefe des Netzes und die Anzahl der Gewichte sowie die Art der Verbindungen machen auch die Komplexität eines KNN aus.

---

[200] Siehe nur *Pasquale*, The black box society; *Martini*, Blackbox Algorithmus – Grundfragen einer Regulierung Künstlicher Intelligenz.

[201] Siehe etwa *Castelvecchi*, Nature 2016, 20.

[202] *Burrell*, Big Data & Society 2016, 1 (1 f.).

[203] „[...] how or why a particular classification has been arrived at from inputs", *Burrell*, Big Data & Society 2016, 1 (1).

[204] *Ebeling/Freund/Schweitzer*, Komplexe Strukturen, S. 13.

[205] *Ebeling/Freund/Schweitzer*, Komplexe Strukturen, S. 24.

[206] *Ribeiro/Singh/Guestrin*, Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining 2016, 1135 (1137); *Guidotti u. a.*, ACM Computing Surveys 2019, 1 (9).

Die strukturelle Komplexität eines KNN macht es praktisch unmöglich, das Netz symbolisch zu beschreiben. Die Beschreibung müsste die gesamte Struktur des Netzes einschließlich der Gewichtungen zwischen den Verbindungen betreffen. Je wahrscheinlicher es ist, das Netz umfassend zu beschreiben, desto geringer ist sein Informationsgehalt. Bei hochkomplexen Strukturen sind die Möglichkeiten der Beschreibung jedoch gering und die (potenzielle) Information des KNN ist groß. Diese Information kann aber nicht notwendigerweise erfasst werden, sodass die faktische Information gering ist.
An dieser Schnittstelle siedelt das Forschungsfeld der XAI an, das versucht, die potenzielle Information u. a. von KNN in faktische Information zu verwandeln, indem die implizit repräsentierte Information explizit gemacht wird. Je besser das gelingt, desto größer ist die faktische Information über das Netz und mithin seine Beschreibbarkeit.
Den verschiedenen Arten von Opazität stehen auch unterschiedliche Arten von Transparenz gegenüber. Nicht jede Art von Transparenz beseitigt dabei alle Formen von Opazität:
Während politische Forderungen nach Transparenz hauptsächlich die erste Form der Opazität (Geheimhaltung) in den Blick nehmen, stellt sich in einem zweiten Schritt für das „Wie" der Transparenz die Frage nach der dritten Form der Opazität (Komplexität). Denn wird Opazität aufgrund von Geheimhaltung durch Transparenzpflichten (vermeintlich) überwunden, so muss anschließend die Opazität aufgrund von Komplexität überwunden werden.
Dabei stellt sich die Frage, was genau an einem Modell opak ist beziehungsweise welche Art von Transparenz gefordert ist. Hier spielt die Art der Wissensrepräsentation in einem System, also die Unterscheidung zwischen explizit und implizit repräsentierter Information eine entscheidende Rolle. Sollen Quellcode und Gewichte eines KNN offengelegt werden, so erübrigt sich Opazität durch Geheimhaltung, jedoch nicht notwendigerweise Opazität durch Komplexität. Denn die Information des Netzes ist in Quellcode und Gewichten nur implizit repräsentiert. Mithilfe von XAI-Techniken kann Opazität durch Komplexität zumindest teilweise überwunden werden, ohne dass Opazität durch Geheimhaltung notwendigerweise entfallen muss.[207]

[207] Dazu unter Kapitel 8.

Je nachdem, wer Adressat einer Transparenzpflicht ist, kommen daher ganz unterschiedliche Erklärungsformen eines KNN beziehungsweise seiner Entscheidungen in Betracht.[208]

B. Warum Transparenz?

Transparenz und Rechenschaftspflicht nehmen in der Debatte um KI-Regulierung eine bedeutende Rolle ein.[209] Besonders im Fokus der Transparenzdebatte stehen sogenannte Algorithmische Entscheidungssysteme (engl. *algorithmic decision making,* ADM), bei denen Menschen beziehungsweise menschliches Verhalten durch ein Verfahren Künstlicher Intelligenz analysiert und bewertet werden und das System anschließend eine Entscheidung fällt oder zumindest vorschlägt.[210]
Eine besonders persönlichkeitsrechtssensible und diskriminierungsträchtige Form eines solchen ADM ist das sogenannte *Profiling*, das in Art. 4 Nr. 4 DSGVO definiert wird als

> „jede Art der automatisierten Verarbeitung personenbezogener Daten, die darin besteht, dass diese personenbezogenen Daten verwendet werden, um bestimmte persönliche Aspekte, die sich auf eine natürliche Person beziehen, zu bewerten, insbesondere um Aspekte bezüglich Arbeitsleistung, wirtschaftliche Lage, Gesundheit, persönliche Vorlieben, Interessen, Zuverlässigkeit, Verhalten, Aufenthaltsort oder Ortswechsel dieser natürlichen Person zu analysieren oder vorherzusagen".

Daran anknüpfend wird noch das sogenannte *Scoring* unterschieden, das in § 31 Abs. 1 BDSG definiert wird als

> „die Verwendung eines Wahrscheinlichkeitswerts über ein bestimmtes zukünftiges Verhalten einer natürlichen Person zum Zweck der Entschei-

---

[208] Siehe dazu *Samek/Müller*, in: Samek u. a., Explainable AI: Interpreting, Explaining and Visualizing Deep Learning, S. 9 f.

[209] *Europäische Kommission*, Weißbuch zur Künstlichen Intelligenz, S. 11; *Hochrangige Expertengruppe für Künstliche Intelligenz*, Ethik-Leitlinien für eine vertrauenswürdige KI, S. 17 f.

[210] *Dreyer/Schulz*, Was bringt die Datenschutz-Grundverordnung für automatisierte Entscheidungssysteme?, S. 9; *Wieder*, in: Taeger, Rechtsfragen digitaler Transformationen - Gestaltung digitaler Veränderungsprozesse durch Recht, S. 507.

> dung über die Begründung, Durchführung oder Beendigung eines Vertragsverhältnisses mit dieser Person".[211]

Der Anwendungsbereich dieser Formen von ADM ist groß: Banken benutzen sie bei der Kreditvergabe, Versicherungen für Tarifangebote, Arbeitgeber bei der Bewerberauswahl, Universitäten bei der Vergabe von Studienplätzen.[212]
Und auch in weniger existenziellen Bereichen wird Künstliche Intelligenz zur Entscheidungsfindung und Differenzierung eingesetzt: Welche Werbung der Internetnutzer im Browser angezeigt bekommt, welche Filme ihm auf Streaming-Webseiten vorgeschlagen werden, welchen Preis seine Flugreise hat – diese kleinen Entscheidungen werden durch Künstliche Intelligenz getroffen und sind individualisiert.[213]
Der Einsatz von ADM zeichnet sich dabei durch eine Informationsasymmetrie zwischen Unternehmern und Verbrauchern aus.[214] Er birgt einige Risiken, etwa für das Persönlichkeitsrecht, die Meinungs- und Informationsfreiheit und den Wettbewerb.[215] Auch Risiken für die Sicherheit und Haftung können durch Künstliche Intelligenz entstehen.
Im Folgenden sollen einige der wichtigsten Risiken beim Einsatz Maschinellen Lernens dargestellt werden, vor deren Hintergrund die Transparenzdebatte zu betrachten ist.
Ein Problem kann sich zu aller erst schlicht dadurch ergeben, dass ein System künstlicher Intelligenz falsche Schlüsse zieht und darauf aufbauend falsche Entscheidungen trifft oder vorschlägt. Denn bei der Datenanalyse sucht Künstliche Intelligenz nach Zusammenhängen zwischen verschiedenen Variablen, den sogenannten Korrelationen. Diese sind jedoch von Kausalzusammenhängen zu unterscheiden, welche die Beziehung zwischen Ursache und Wirkung beschreiben. Die Verwechslung von Korrelation und Kausalität durch KI kann zu grotesken

---

[211] Scoring ist somit eine Form des Profilings, vgl. *Buchner*, in: Kühling/Buchner DS-GVO, Art. 22 Rn. 22.

[212] *Martini*, Blackbox Algorithmus – Grundfragen einer Regulierung Künstlicher Intelligenz, S. 47 und 53 f. m. w. N.

[213] Vgl. zu personalisierten Preisen *Golland*, in: Taeger, Die Macht der Daten und der Algorithmen : Regulierung von IT, IoT und KI, S. 61 ff.; *Martini*, Blackbox Algorithmus – Grundfragen einer Regulierung Künstlicher Intelligenz, S. 52 f.

[214] Näheres bei *Gesellschaft für Informatik*, Technische und rechtliche Betrachtungen algorithmischer Entscheidungsverfahren, S. 75 ff.

[215] *Martini*, Blackbox Algorithmus – Grundfragen einer Regulierung Künstlicher Intelligenz, S. 29 ff.

Ergebnissen führen, die ein rational denkender Mensch ohne weiteres erkennt.[216] Die Maschine jedoch ist zu einer solchen Bewertung ihrer Ergebnisse nicht fähig und könnte aus entsprechenden Daten einen Kausalzusammenhang und darauf aufbauend eine Entscheidung oder Handlungsempfehlung ableiten.
Darüber hinaus birgt der Einsatz künstlicher Intelligenz ein großes Diskriminierungspotenzial.[217] Das mag zunächst verwundern, denn die Entscheidung durch die Maschine erweckt auf den ersten Blick den Anschein von Objektivität, der „Risikofaktor Mensch" scheint überwunden.[218] Der Bankkunde, der einen Kredit benötigt, sieht sich nicht einem möglicherweise müden, schlecht gelaunten, erkälteten oder vorurteilsbehafteten Mitarbeiter gegenüber, der persönliche Befindlichkeiten in seine Entscheidung einfließen lässt.
Allerdings fließen auch in ein KI-Modell Wissen und Wertentscheidungen des Programmierers ein.[219] Zudem wird KI mit Daten trainiert und ist dadurch geneigt, gesellschaftliche Missstände zu reproduzieren. Denn der Einzelne wird aufgrund bestimmter Merkmale und Präferenzen einer Gruppe zugeordnet und nicht als Individuum bewertet.
Eine anhand dieser Gruppenzuordnung getroffene Entscheidung kann weitreichende Konsequenzen für den gesamten Lebensweg haben.[220] Dies veranschaulicht ein prominentes Beispiel aus den USA: in vielen US-Bundesstaaten wird bei Gericht zur Prognose von Rückfallwahrscheinlichkeiten von Straftäterinnen und Straftätern die Software COMPAS eingesetzt. Diese wurde anhand reeller Daten trainiert und soll die Entscheidung über Strafmaß und Bewährung unterstützen. Auswertungen des Prognoseverfahrens deuten jedoch auf eine Diskriminierung von *People of Colour* durch das System hin.[221] Denn die Anzahl derer,

---

[216] So korrelieren beispielsweise in den USA zwischen 2000 und 2009 der Käseverzehr pro Kopf mit der Anzahl der Personen, die dadurch gestorben sind, dass sie sich in ihrem Bettzeug verheddert haben. Siehe für dieses und weitere, nicht weniger kuriose Beispiele *Vigen*, Spurious correlations, https://www.tylervigen.com/spurious-correlations (zuletzt abgerufen am 26.10.2023).

[217] Dazu ausführlich *Gesellschaft für Informatik*, Technische und rechtliche Betrachtungen algorithmischer Entscheidungsverfahren, S. 84 ff.

[218] Vgl. dazu *Martini*, Blackbox Algorithmus – Grundfragen einer Regulierung Künstlicher Intelligenz, S. 47 f.

[219] *Martini*, Blackbox Algorithmus – Grundfragen einer Regulierung Künstlicher Intelligenz, S. 335.

[220] Vgl. dazu *Martini*, Blackbox Algorithmus – Grundfragen einer Regulierung Künstlicher Intelligenz, S. 49 ff.

[221] *Angwin/Larson/Mattu/Kirchner*, Machine Bias, ProPublica, https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing (zu-

denen eine hohe Rückfallwahrscheinlichkeit vorhergesagt wurde und die nicht rückfällig wurden (Falsch-Positive), ist bei dunkelhäutigen fast doppelt so hoch wie bei weißen Angeklagten. Dies erstaunt umso mehr, als die Ethnie kein Merkmal ist, mit dem das System trainiert wurde.[222]

Darin zeigt sich ein weiteres Problem des Einsatzes künstlicher Intelligenz. Um nämlich der Verfestigung struktureller Diskriminierung entgegenzuwirken, werden Trainingsdatensätze häufig bereinigt oder ausbalanciert, oder es werden andere Methoden, etwa die *differential privacy*, angewendet, die eine Bewertung anhand diskriminierungsträchtiger Merkmale ausschließen sollen.[223] Doch auch nach einem Training mit bereinigten (*unbiased*) Datensätzen ergibt sich das gerade beschriebene Problem. Denn Künstliche Intelligenz, und insbesondere ein Künstliches Neuronales Netz, kann auch weniger stark ausgeprägte Abhängigkeiten zwischen einzelnen Variablen ausfindig machen.[224] Dann besteht die Gefahr, dass Diskriminierungen aufgrund von mit der Hautfarbe korrelierenden Merkmalen (Stellvertretervariablen, sog. *proxys*) doch wieder Einfluss in die Entscheidungsfindung erhalten und auf diese Weise perpetuiert werden.[225]

Die vorangegangenen Beispiele machen die Grundrechtsrelevanz eines Einsatzes von KI deutlich.[226] Auch gesamtgesellschaftlich gesehen können durch Steuerung von Meinungsbildern und Wahlbeeinflussung in sozialen Netzwerken

---

letzt abgerufen am 26.10.2023).; *Martini*, Blackbox Algorithmus – Grundfragen einer Regulierung Künstlicher Intelligenz, S. 55 ff.; der EU-Gesetzgeber bewertet den Einsatz von KI durch Strafverfolgungsbehörden und Justiz überwiegend als „hochriskant", Entwurf der KI-Verordnung, Allgemeine Ausrichtung, Art. 6 Abs. 2 i. V. m. Anhang III Nr. 6, 8; siehe auch Erwägungsgrund 38.

[222] *Martini*, Blackbox Algorithmus – Grundfragen einer Regulierung Künstlicher Intelligenz, S. 55 ff.

[223] *Martini*, Blackbox Algorithmus – Grundfragen einer Regulierung Künstlicher Intelligenz, S. 243; *Gesellschaft für Informatik*, Technische und rechtliche Betrachtungen algorithmischer Entscheidungsverfahren, S. 35; vgl. zu Anforderungen an Trainings-, Validierungs- und Testdatensätze Entwurf der KI-Verordnung, Allgemeine Ausrichtung, Art. 10.

[224] *Martini*, Blackbox Algorithmus – Grundfragen einer Regulierung Künstlicher Intelligenz, S. 24.

[225] Vgl. zu diesen Schwierigkeiten der Datenbereinigung *Martini*, Blackbox Algorithmus – Grundfragen einer Regulierung Künstlicher Intelligenz, S. 55 ff.; *Gesellschaft für Informatik*, Technische und rechtliche Betrachtungen algorithmischer Entscheidungsverfahren, S. 36 f.

[226] Siehe zu den Grundrechten, die durch KI-Regulierung geschützt werden sollen Entwurf der KI-Verordnung, Allgemeine Ausrichtung, S. 12 f.

(sog. *Microtargeting*) Risiken für die freiheitlich-demokratische Grundordnung entstehen.[227]

Die oben beschriebenen technischen Risiken können auch Risiken für die Sicherheit von Anwendern generieren. Fällt eine medizinische KI-Anwendung beispielsweise eine falsche Diagnose, die auch dem menschlichen Anwender nicht auffällt, kann das weitreichende Konsequenzen für Leib und Leben des Betroffenen haben. Gleiches gilt etwa für die Anwendung von KI im autonomen Fahren. An diese Sicherheitsproblematik können sich Haftungsrisiken anschließen, die sich aus der Opazität vieler KI-Anwendungen und der Vielzahl ungeklärter Haftungsfragen ergeben.[228]

Die Notwendigkeit von Transparenz und Regulierung von ML-Systemen ergibt sich jedoch auch vom entgegengesetzten Standpunkt aus. Bisher standen die Nachteile und Risiken von KI im Fokus, aber ihr Einsatz zur Entscheidungsfindung hat offensichtlich auch enorme Vorteile. Dies lässt sich wiederum gut am Beispiel medizinischer Diagnostik verdeutlichen. Ein System künstlicher Intelligenz kann anhand eines Datensatzes trainiert werden, der den Erfahrungsschatz eines menschlichen Experten weit übersteigt. Anhand dieser Daten kann es Zusammenhänge erkennen, die dem menschlichen Betrachter verborgen bleiben.

Der Einsatz kann somit von großem Vorteil sein, sofern es gelingt, die beschriebenen Risiken zu minimieren. In vielen Aspekten ist eine Entscheidung durch künstliche Intelligenz der menschlichen überlegen, weshalb auch reflektiert werden muss, wann ein Recht auf eine Entscheidung durch Künstliche Intelligenz beziehungsweise eine Pflicht zu deren Nutzung bestehen sollte.[229] Gerade aus haftungsrechtlicher Sicht wird sich diese Frage zunehmend stellen.[230]

Gleichzeitig entfaltet der Einsatz einer KI allein schon durch die schiere Masse an Fällen, die sie in kürzester Zeit erfassen kann, eine weitaus größere Wirkmacht als ein menschlicher Entscheidungsträger. Auch diese „Breitenwirkung" muss bei der Frage der Regulierung berücksichtigt werden.[231]

---

[227] Vgl. zu diesem „Wandel der Medienordnung" *Martini*, Blackbox Algorithmus – Grundfragen einer Regulierung Künstlicher Intelligenz, S. 99 ff.

[228] *Europäische Kommission*, Weißbuch zur Künstlichen Intelligenz, S. 15.

[229] Siehe zum Ganzen: *Martini*, Blackbox Algorithmus – Grundfragen einer Regulierung Künstlicher Intelligenz, S. 47 f.; *Wischmeyer*, AöR 2018, 1 (33) Fn. 128 m. w. N.

[230] Siehe dazu etwa *Hacker/Krestel/Grundmann/Naumann*, Artificial Intelligence and Law 2020, 415 (416).

[231] *Martini*, Blackbox Algorithmus – Grundfragen einer Regulierung Künstlicher Intelligenz, S. 89.

Dieser kleine Ausschnitt aus der Transparenzdebatte macht deutlich, dass der Einsatz von KI vielfältige Auswirkungen auf die Grundrechte, den freiheitlichen Rechtsstaat und die Sicherheit haben kann. Vielfach wird angenommen, dass das geltende Recht nicht die notwendigen Instrumente bietet, diese Werte ausreichend zu schützen.[232]

Der Ruf nach mehr Transparenz von KI-Systemen hat daher durchaus seine Berechtigung. Denn bei den genannten Beispielen kommen teilweise klassische Algorithmen zum Einsatz, die noch verhältnismäßig einfach zu verstehen sind, zumindest von Experten.[233] Teilweise werden jedoch auch komplexere Systeme Künstlicher Intelligenz, wie etwa Künstliche Neuronale Netze und generell *Deep Learning*, eingesetzt, die auch für Fachleute in großen Teilen unergründlich sind. Gleichzeitig sind allerdings auch die Rechte der betroffenen Unternehmen über Art. 16 beziehungsweise Art. 17. GRCh geschützt.[234] Daher wird im folgenden Teil auch untersucht, inwieweit Regelwerke neben Transparenzpflichten auch Vorschriften zum Schutz von Geschäftsgeheimnissen vorsehen.

C. Panorama der Transparenzregulierung und Rolle des Geheimnisschutzes

Die Informations- und Auskunftsrechte der DSGVO stehen seit geraumer Zeit im Fokus der Frage nach Transparenz von KI-Systemen und sollen daher auch hier den Ausgangspunkt der Untersuchung darstellen. Mittlerweile steht jedoch mit dem Entwurf der KI-Verordnung ein erster Versuch der Regulierung Künstlicher Intelligenz auf EU-Ebene bereit, der ebenfalls Transparenzpflichten für KI-Systeme vorsieht. Insgesamt finden Regelungen zur Transparenz von

---

[232] *Gesellschaft für Informatik*, Technische und rechtliche Betrachtungen algorithmischer Entscheidungsverfahren, S. 7: „erhebliche Defizite im geltenden Recht"; *Martini*, Blackbox Algorithmus – Grundfragen einer Regulierung Künstlicher Intelligenz, S. 109.

[233] Beim Kredit-Scoring werden etwa immer noch am häufigsten statistische Verfahren wie die lineare und logistische Regression eingesetzt, *Thomas/Edelman/Crook*, in: Credit scoring and its applications, Chapter 3, S. 25; *Gesellschaft für Informatik*, Technische und rechtliche Betrachtungen algorithmischer Entscheidungsverfahren, S. 32.

[234] Siehe nur *Kalbfus*, in: Harte-Bavendamm/Ohly/Kalbfus, GeschGehG, Einl. A Rn. 165; *Alexander*, MMR 2021, 690 (691).

KI-Systemen vermehrt Eingang in EU-Rechtsakte.[235] Eine Analyse der umfangreichen Regulierung ist im Rahmen dieser Arbeit nicht möglich, allerdings auch nicht notwendig. Denn die grundlegenden Fragen nach Art und Empfänger der Information lassen sich auch von bestehenden Normen abstrahieren und können so auch auf weitere (zukünftige) Regelungen angewandt werden. Ziel dieses Teils ist daher nicht die tiefgehende juristische Analyse der Regulierungslandschaft, sondern die Analyse der Grundzüge der verschiedenen Regelungen, also namentlich der Frage, was gegenüber wem offenbart werden muss. Außerdem wird untersucht, ob und in welchem Umfang die betreffenden Regelungen den Schutz von Geschäftsgeheimnissen vorsehen.

Anhand der Regelungen und Regelungsvorschläge aus DSGVO und KI-VO-E soll ein, wenn auch unvollständiges, Panorama der Transparenzregulierung auf EU-Ebene gezeichnet werden, vor dessen Hintergrund das Konfliktfeld von Transparenz und Offenbarung von Geschäftsgeheimnissen anschaulich diskutiert werden kann.

Die Vielzahl möglicher Regulierungsinstrumente wird meist nach dem Zeitpunkt einer möglichen Intervention gegliedert. Es können dann grob unterschieden werden: die abstrakte Erklärung eines Entscheidungssystems von der konkreten Erklärung einer einzelnen Entscheidung sowie eine der Entscheidung vorgelagerte Erklärung (ex ante) von einer der Entscheidung nachfolgende Erklärung (ex post).[236]

Dieser Darstellung wird auch hier gefolgt. Dabei wird jedoch nur auf diejenigen Regelungen eingegangen, welche die (geheime) Information eines trainierten KNN gefährden könnten, weil sie auf Transparenz abzielen. Nur diese Art von Regulierung ist für die vorliegende Fragestellung relevant und sie ist gleichzeitig besonders wesentlich für den rechtlichen Schutz Betroffener gegenüber ADM.[237]

---

[235] Siehe etwa Art. 5 Abs. 1 und Abs. 2 VO (EU) 2019/1159 (P2B-Verordnung), Art. 7 Abs. 4a RL 2005/29/EG (UGP-Richtlinie) und Art. 6a Abs. 1 RL 2011/83/EU (VRRL). Siehe für eine Analyse dieser Normen aus der Sicht des Geheimnisschutzrechts *Alexander*, MMR 2021, 690.

[236] Siehe dazu nur *Wachter/Mittelstadt/Floridi*, IDPL 2017, 76 (78).

[237] *Gesellschaft für Informatik*, Technische und rechtliche Betrachtungen algorithmischer Entscheidungsverfahren, S. 161.

I. Datenschutzgrundverordnung[238]

1. Informationspflicht (Artikel 13 und 14 DSGVO)

Informationspflichten über die abstrakte Entscheidungsfindung (ex ante) sind in der DSGVO in Art. 13 Abs. 2 lit. f und Art. 14 Abs. 2 lit. g normiert.[239] Danach muss der Verantwortliche bei der Erhebung personenbezogener Daten der betroffenen Person zum Zeitpunkt der Erhebung der Daten unter anderem Informationen zur Verfügung stellen über:

> „das Bestehen einer automatisierten Entscheidungsfindung einschließlich Profiling gemäß Artikel 22 Absatz 1 und Absatz 4 und – zumindest in diesen Fällen – aussagekräftige Informationen über die involvierte Logik sowie die Tragweite und die angestrebten Auswirkungen einer derartigen Verarbeitung für die betroffene Person." (Art. 13 Abs. 2 lit. f DSGVO).

Eine Parallelregelung für den Fall, dass die personenbezogenen Daten nicht bei der betroffenen Person selbst erhoben werden, ist in § 14 Abs. 2 lit. g DSGVO getroffen.
Beide Normen sind durch den Verweis auf Art. 22 Abs. 1 DSGVO in ihrem Anwendungsbereich beschränkt. Nach Art. 22 Abs. 1 DSGVO hat die betroffene Person das Recht, „nicht einer ausschließlich auf einer automatisierten Verarbeitung – einschließlich Profiling – beruhenden Entscheidung unterworfen zu werden, die ihr gegenüber rechtlicher Wirkung entfaltet oder sie in ähnlicher Weise erheblich beeinträchtigt."
Nicht unter Art. 22 und mithin auch nicht unter die Art. 13 und 14 DSGVO fallen daher zum einen die häufigen Fälle, in denen KI lediglich zur Vorbereitung einer menschlichen Entscheidung eingesetzt wird, da die Entscheidung dann nicht „ausschließlich" auf einer automatisierten Verarbeitung beruht. Nach aktuellem Recht muss demnach grundsätzlich nur bei vollautomatisierten Verfahren offengelegt werden, dass für die Entscheidungsfindung Künstliche

[238] Ein Auszug aus der DSGVO mit den hier analysierten Normen findet sich im Anhang.
[239] Siehe zu dieser Ausrichtung des Informationsanspruchs nur *Wachter/Mittelstadt/Floridi*, IDPL 2017, 76; *Veil*, in: Gierschmann DSGVO, Art. 13 und 14 Rn. 120.

Intelligenz eingesetzt wird („Ob" des Einsatzes).[240] Zum anderen fallen diejenigen Fälle nicht unter die Informationspflicht, in denen die Entscheidung keine rechtliche Wirkung oder ähnliche erhebliche Beeinträchtigung für den Betroffenen mit sich bringt.

Wann aufgrund der in Spiegelstrichen eingefügten Einschränkung „zumindest in diesen Fällen" auch aussagekräftige Informationen über die involvierte Logik etc. zur Verfügung gestellt werden muss, ist ebenfalls strittig.

Teilweise wird angenommen, dass die erweiterte Informationspflicht über die involvierte Logik auch bei Entscheidungen Anwendung finden soll, die nicht unter Art. 22 Abs. 1 und 4 DSGVO fallen, da sie nicht vollautomatisiert ablaufen oder keine rechtliche Relevanz für den Betroffenen haben.[241] Andere lehnen dies ab und nehmen eine Informationspflicht über das „Wie" der Entscheidung nur bei vollautomatisierten Entscheidungen aufgrund von Profiling oder bei ähnlicher persönlichkeitsrechtlicher Sensibilität an.[242]

Zu guter Letzt ist unklar, was genau „Informationen über die involvierte Logik" sind und wie der Informationspflicht nachgekommen werden kann.

Die Debatte wird vor allem im Hinblick auf die Frage intensiv geführt, inwiefern „Algorithmen" offengelegt werden müssen.[243] Dabei arbeitet sich die Diskussion augenscheinlich an zwei Variablen ab, die (noch) nicht eindeutig bestimmt sind:

Einerseits ist unklar, welche Bedeutung dem Schutz von Geschäftsgeheimnissen zukommen und wie er dogmatisch gewährleistet sein soll. Andererseits soll die zur Verfügung gestellte Information „in präziser, transparenter, verständlicher

---

[240] Vgl. zu den Einzelheiten *Martini*, Blackbox Algorithmus – Grundfragen einer Regulierung Künstlicher Intelligenz, S. 169 ff.; *Hennemann*, in: Paal/Pauly DS-GVO, Art. 13 Rn. 31; *Kumkar/Roth-Isigkeit*, JZ 2020, 277; andere vertreten, dass im Falle von Profiling auch bei bloßer Entscheidungsunterstützung eine einfache Informationspflicht bestehen soll, *Buchner*, in: Kühling/Buchner DS-GVO, Art. 13 Rn. 52; *Mester*, in: Taeger/Gabel DS-GVO, Art. 13 Rn. 27.

[241] *Mester*, in: Taeger/Gabel DS-GVO, Art. 13 Rn. 28; *Franck*, in: Gola DS-GVO, Art. 13 Rn. 27 m.w.N. *Buchner*, in: Kühling/Buchner DS-GVO, Art. 13 Rn. 53 m.w.N.

[242] *Martini*, Blackbox Algorithmus – Grundfragen einer Regulierung Künstlicher Intelligenz, S. 185 mit vorangehender sehr ausführlicher Analyse der Normen.

[243] Der Begriff wird h.E. in der Debatte über „Algorithmen-Regulierung" zu undifferenziert, gewissermaßen als Oberbegriff, gebraucht. Bei den besonders opaken und daher in Bezug auf Transparenz gerade relevanten KNN handelt es sich jedoch nicht um Algorithmen. Die Netze werden lediglich durch einen (Lern-)Algorithmus trainiert, die Arbeitsweise eines trainierten KNN ist dann gerade nicht deterministisch wie ein Algorithmus, sondern probabilistisch.

und leicht zugänglicher Form in einer klaren und einfachen Sprache“ übermittelt werden (Art. 12 Abs. 1 S. 1 DSGVO), was zu technische Darstellungsformen – Geschäftsgeheimnisse hin oder her – ausschließen dürfte.
In diesen beiden Variablen – Geheimnisschutz und Verständlichkeitsgebot – spiegelt sich die bereits eingeführte Dichotomie von Opazität durch Geheimhaltung und Opazität durch Komplexität wider.
Die DSGVO enthält keine eindeutige, allgemeine Regelung zum Ausgleich der Betroffenenrechte mit den Rechten des Verantwortlichen, insbesondere nicht im verfügenden Teil. In Bezug auf Geschäftsgeheimnisse kann daher lediglich auf den Erwägungsgrund 63 verwiesen werden, der sich jedoch ausdrücklich nur auf das in Art. 15 DSGVO normierte Auskunftsrecht bezieht.[244]
Häufig wird daher hinsichtlich des Informationsumfangs der Art. 13 und 14 DSGVO schlicht ohne nähere Begründung auf eine Beschränkung durch Geschäftsgeheimnisse verwiesen.[245] Teilweise wird für eine Lösung durch eine beschränkende Regelung nach Maßgabe des Art. 23 DSGVO oder für eine über den Verweis auf Treu und Glauben in Art. 5 Abs. 1 lit. a DSGVO hinausgehende „allgemeine Rechtsgüterabwägung“ plädiert.[246]
Abgesehen von der dogmatischen Anknüpfung des Schutzes von Geschäftsgeheimnissen sind auch die Parameter unklar, nach denen die wohl erforderliche Abwägungsentscheidung getroffen werden soll.[247]
Auch wenn noch zum alten Recht ergangen, wird das Urteil des BGH zur Scoreformel der SCHUFA häufig als Maßstab angeführt.[248] Der BGH verneint dort die von der Klägerin begehrte Offenlegung der Scoreformel der SCHUFA, also deren abstrakte Rechenformel zur Berechnung des Scorewerts ihrer Kunden.

---

[244] Siehe dazu unten.

[245] Siehe etwa *Paal/Hennemann*, in: Paal/Pauly DS-GVO, Art. 13 Rn. 31b f.; über eine teleologische Reduktion zur Wahrung der Rechte des Verantwortlichen *Veil*, in: Gierschmann DSGVO, Art. 13 Rn. 121.

[246] Zu Art. 23 *Bäcker*, in: Kühling/Buchner DS-GVO, Art. 13 Rn. 54; zu Art. 5 Abs. 1 lit. a *Franck*, in: Gola DS-GVO, Art. 13 Rn. 28.

[247] Für eine enge Auslegung der Schutzwürdigkeit von Geschäftsgeheimnissen *Franck*, in: Gola DS-GVO, Art. 15 Rn. 34.

[248] BGH, Urteil v. 28.1.2014, VI ZR 156/13, NJW 2014, 1235. Gegen das Urteil ist eine Verfassungsbeschwerde anhängig, Az. 1 BvR 756/2014. Das Auskunftsverlangen wurde unter anderem auf § 34 Abs. 4 S. 1 Nr. 4 BDSG a.F. gestützt, nach dem über „das Zustandekommen und die Bedeutung der Wahrscheinlichkeitswerte einzelfallbezogen und nachvollziehbar in allgemein verständlicher Form“ Auskunft erteilt werden musste.

Das Gericht schließt dabei explizit die Algorithmen der Scoreformel, die Gewichtungen einzelner Parameter und Vergleichsgruppen von der Offenlegungspflicht aus. In seiner Entscheidung räumt der BGH somit dem Schutz von Geschäftsgeheimnissen Vorrang vor der vollen Transparenz des Entscheidungsmechanismus ein.

Vertreter einer restriktiven Auslegung halten entsprechend die Offenlegung der generellen Funktionsweise durch Beschreibung etwa von Parametern und der Grundstruktur des Entscheidungsmechanismus für ausreichend, während „der Algorithmus" nicht offenbart werden soll.[249] Andere favorisieren auch eine Offenlegung von „Algorithmen" oder von Gewichtungen der Parameter.[250] Sie verweisen darauf, dass eine „pauschale Bevorzugung" der Interessen von Unternehmen gegenüber derer von Betroffenen „der DSGVO fremd" sei,[251] oder dass die DSGVO sogar das Gewicht in Richtung der Auskunftsrechte verlagert habe.[252]

Eine Offenlegung des Quellcodes des verwendeten Programms wird jedoch überwiegend als unvereinbar mit dem Schutz von Geschäftsgeheimnissen und als nicht zielführend abgelehnt.[253]

Bei der Diskussion um die Reichweite der „involvierten Logik" spielt neben dem Schutz von Geschäftsgeheimnissen noch die erwähnte zweite unbestimmte Variable eine Rolle: die zur Verfügung gestellte Information muss für den Betroffe-

---

[249] *Martini*, Blackbox Algorithmus – Grundfragen einer Regulierung Künstlicher Intelligenz, S. 181 f.; *Paal/Hennemann*, in: Paal/Pauly DS-GVO, Art. 13 Rn. 31c; *Kumkar/Roth-Isigkeit*, JZ 2020, 277 (285); *Veil*, in: Gierschmann DSGVO, Art. 13 und 14 Rn. 128; die Entscheidung müsse „nachvollziehbar", nicht „nachrechenbar" sein: OLG Nürnberg, Urteil v. 20.10.2012, 3 U 2362/11, ZD 2013, 26 (27) – *Auskunftsanspruch über Scorewert*, zum BDSG a. F..

[250] Für die Herausgabe des „Algorithmus" etwa *Roßnagel/Nebel/Richter*, ZD 2015, 455 (458); *Mester*, in: Taeger/Gabel DS-GVO, Art. 13 Rn. 29; auch Franck, der dazu die Herausgabe weiterer „erklärender Bestandteile" fordert, „da der durchschnittliche Betroffene mit schematischen Entscheidungsbäumen oder gar Source Code überfordert sein wird", *Franck*, in: Gola DS-GVO, Art. 13 Rn. 26; zu Gewichtungen vgl. *Kaminski*, Berkely Tech L. J. 2019, 190 (214); *Dix*, in: Simitis/Hornung/Spiecker Datenschutzrecht, Art. 13 Rn. 17.

[251] *Franck*, in: Gola DS-GVO, Art. 13 Rn. 28.

[252] *Selbst/Powles*, IDPL 2017, 233 (242); Malgieri/Commandé sprechen von einer „preference for Human Rights of individuals": *Malgieri/Comandé*, IDPL 2017, 243 (262 ff.).

[253] *Martini*, Blackbox Algorithmus – Grundfragen einer Regulierung Künstlicher Intelligenz, S. 181; siehe zur Code-Analyse auch *Gesellschaft für Informatik*, Technische und rechtliche Betrachtungen algorithmischer Entscheidungsverfahren, S. 141 f.

nen verständlich sein (Art. 12 Abs. 1 DSGVO).[254] Vor diesem Hintergrund erscheint es kaum zielführend, dem Betroffenen allzu detaillierte technische Informationen zur Verfügung zu stellen.[255] Hier kommen daher die unter A. dargestellten Formen von Opazität zum Tragen: die Information muss so ausgestaltet sein, dass die Opazität durch Komplexität (3. Stufe) und gegebenenfalls die Opazität mangels Programmierkenntnisse (2. Stufe) überwunden werden.
Das Erfordernis der Verständlichkeit kann dann möglicherweise indirekt zu einer Wahrung von Geschäftsgeheimnissen führen – denn die Opazität durch Geheimhaltung wird durch die „verständliche" Information nicht zwingend beseitigt. Dennoch sollte die Frage der Verständlichkeit zunächst klar von der Frage des Geheimnisschutzes getrennt werden.

2. Auskunftsrecht (Artikel 15 DSGVO)

Nach erfolgtem Einsatz von Künstlicher Intelligenz zur Entscheidungsfindung stellt sich die Frage, inwiefern die getroffene Entscheidung gegenüber der betroffenen Person (ex post) begründet werden muss. Es geht somit nicht mehr um die abstrakte Erklärung eines Entscheidungsmechanismus, sondern um die nachträgliche Begründung eines konkreten Ergebnisses, also etwa der Einstufung eines potenziellen Kunden als kreditwürdig oder kreditunwürdig durch eine Bank (sog. Recht auf Erklärung).
Ob das in Art. 15 Abs. 1 lit. h DSGVO als Parallelvorschrift zu den Art. 13 Abs. 2 lit. f und Art. 14 Abs. 2 lit. g DSGVO gestaltete Auskunftsrecht der betroffenen Person eine Begründungspflicht normiert, ist höchst strittig. Teilweise wird ein Recht auf Erklärung (*right to explanation*) der betroffenen Person und mithin ein Recht auf eine ex-post-Erklärung einer konkreten Entscheidung angenommen.[256] Andere sehen auch in Art. 15 Abs. 1 lit. h DSGVO nur ein Auskunftsrecht über die abstrakte Funktionsweise der Entscheidungsfindung.[257]

---

[254] Siehe dazu nur *Artikel-29-Datenschutzgruppe*, Leitlinien für Transparenz gemäß der Verordnung 2016/679, WP 260 rev.01, S. 7 ff.; *Malgieri/Comandé*, IDPL 2017, 243.

[255] So auch zur „Effektivität eines ‚Rechts auf Erklärung'" *Wischmeyer*, AöR 2018, 1 (52 ff.).

[256] Siehe nur *Malgieri/Comandé*, IDPL 2017, 243 (255 f.); *Edwards/Veale*, Duke Law & Technology Review 2017, 18 (52); *Dix*, in: Simitis/Hornung/Spiecker Datenschutzrecht, Art. 15 Rn. 25; *Franck*, in: Gola DS-GVO, Art. 15 Rn. 19 m. w. N.

[257] Siehe z. B. *Wachter/Mittelstadt/Floridi*, IDPL 2017, 76 (83 ff.); *Wischmeyer*, AöR 2018, 1 (50 ff.); *Martini*, Blackbox Algorithmus – Grundfragen einer Regulierung Künstlicher Intelligenz, S. 190 ff.

Für die Art der zur Verfügung zu stellenden Information gelten die zu den Art. 13 und 14 DSGVO getätigten Ausführungen entsprechend: sie könnte „Algorithmen", Parameter oder auch deren Gewichtung umfassen.
Auch Art. 15 Abs. 1 lit. h DSGVO ist zudem grundsätzlich auf rein automatische Entscheidungsfindung und erhebliche Beeinträchtigung der betroffenen Person begrenzt (vgl. Art. 22 Abs. 1 DSGVO).[258]
Hinsichtlich des Schutzes von Geschäftsgeheimnissen besteht bei Art. 15 DSGVO die Besonderheit, dass nach Absatz 4 das „Recht auf Erhalt einer Kopie gemäß Absatz 3 [...] die Rechte und Freiheiten anderer Personen nicht beeinträchtigen [darf]." Geschäftsgeheimnisse werden unter Verweis auf Erwägungsgrund 63 als Rechte anderer Personen und daher berücksichtigungswert eingeordnet. In Erwägungsgrund 63 heißt es:

> „Dieses Recht [auf Auskunft, *Anm. der Verfasserin*] sollte die Rechte und Freiheiten anderer Personen, etwa Geschäftsgeheimnisse oder Rechte des geistigen Eigentums und insbesondere das Urheberrecht an Software, nicht beeinträchtigen. Dies darf jedoch nicht dazu führen, dass der betroffenen Person jegliche Auskunft verweigert wird."

Dieser Passus umreißt das beschriebene „Spannungsverhältnis"[259] zwischen Auskunftsrecht und Geschäftsgeheimnissen, ohne jedoch eine Möglichkeit zur Auflösung der widerstreitenden Interessen zu benennen.
Es dürfte zwar fraglich sein, ob die Kopie gemäß Art. 15 Abs. 3 DSGVO überhaupt die Information über die involvierte Logik gemäß Art. 15 Abs. 1 lit. h umfasst.[260] Dennoch wird überwiegend angenommen, dass Art. 15 Abs. 4 DSGVO allgemein die Auskunftspflicht des Art. 15 Abs. 1 DSGVO einschränkt.[261]
Dementsprechend stellt sich die dogmatische Anknüpfung des Schutzes von Geschäftsgeheimnissen bei Art. 15 DSGVO eindeutiger dar als hinsichtlich der Art. 13 und 14 DSGVO. Einen klar abgesteckten Rahmen für die danach erforderliche Abwägung der widerstreitenden Interessen des Betroffenen und des Verantwortlichen lässt sich der Norm jedoch ebenso wenig entnehmen.

---

[258] a. A. *Edwards/Veale*, Duke Law & Technology Review 2017, 18 (53).
[259] *Mester*, in: Taeger/Gabel DS-GVO, Art. 13 Rn. 29.
[260] Ablehnend *Franck*, in: Gola DS-GVO, Art. 15 Rn. 33.
[261] *Specht*, in: Sydow DSGVO, Art. 15 Rn. 22; *Dix*, in: Simitis/Hornung/Spiecker Datenschutzrecht, Art. 15 Rn. 34; *Paal*, in: Paal/Pauly DS-GVO, Art. 15 Rn. 41.

Auch für die Frage der Verständlichkeit der zur Verfügung gestellten Information gelten die bereits oben dargestellten Unsicherheiten entsprechend.

3. Fazit

Dieser kurze Überblick über die im Datenschutzrecht geführte Diskussion um Transparenz von automatisierten Entscheidungssystemen macht bereits die große Unsicherheit deutlich, die an der Schnittstelle zwischen Datenschutzrecht, Geheimnisschutz und Informationstechnologie liegt. Alle drei Disziplinen sind in der Transparenzfrage eng miteinander verflochten und es bedarf einer genauen und differenzierten Analyse, um die jeweiligen Stränge entwirren und bewerten zu können. Denn die Verpflichtungen aus der DSGVO sind durch den Schutz von Geschäftsgeheimnissen begrenzt. Wie weit diese Begrenzung geht, ist keine dogmatische Frage, die abstrakt von der jeweils eingesetzten Technik beantwortet werden könnte. Ihre Beantwortung ist vielmehr maßgeblich davon abhängig, was überhaupt das Geschäftsgeheimnis eines automatischen Entscheidungssystems ausmacht und in welchen Darstellungsformen es offenbart wird. Zur Beantwortung dieser Frage bedarf es wiederum eines guten Verständnisses der jeweils eingesetzten Technik.

Vor dem Hintergrund dieser Gemengelage erklärt sich auch, warum die Analysen der Informations- und Auskunftsrechte in Bezug auf die „involvierte Logik" automatisierter Entscheidungsfindung, von einzelnen Nennungen konkreter Darstellungsmöglichkeiten abgesehen,[262] überwiegend technikneutral und relativ vage bleiben. Und solange die Phrase „Informationen über die involvierte Logik" nicht (durch die Gerichte) mit konkretem, technikspezifischem Inhalt gefüllt wird, wird sich daran wohl kaum etwas ändern.

In den folgenden Kapiteln wird daher der Versuch unternommen, die Fäden der drei Disziplinen zu entwirren, indem der Fokus vom Datenschutzrecht auf das Geschäftsgeheimnisrecht und die Darstellung von Information verlagert wird. Kann auf diese Weise analysiert werden, was genau die Geschäftsgeheimnisse sind, die einer Auskunft an den Betroffenen im Wege stehen könnten, und ob sie durch die Art der Transparenz überhaupt offengelegt werden, so erübrigt sich dadurch gegebenenfalls ein großer Teil der skizzierten Debatte.

Zunächst soll jedoch noch der aktuell wohl wichtigste Beitrag zur Regulierung von Künstlicher Intelligenz auf seine Transparenzpflichten hin untersucht werden.

---

[262] *Edwards/Veale*, Duke Law & Technology Review 2017, 18 (55 ff.).

II. Entwurf der KI-Verordnung[263]

Das Datenschutzrecht stößt hinsichtlich der Transparenz von Künstlicher Intelligenz schnell an seine Grenzen: zum einen sehen DSGVO und BDSG nur relativ begrenzte Regelungsmöglichkeiten für Künstliche Intelligenz und insbesondere ADM vor, deren Ausmaß darüber hinaus teilweise auch noch umstritten und noch nicht höchstrichterlich geklärt ist.
Zum anderen können durch eine datenschutzkonforme Datenverarbeitung allein Nichtdiskriminierung und gesellschaftliche Teilhabe nicht erreicht werden.[264] Denn der klassische Ansatzpunkt des Datenschutzrechts passt nicht mehr zu den Risiken, die sich aus der Datenverarbeitung durch KI-Systeme ergeben: die Gefahr ergibt sich nicht aus der Verarbeitung eines einzelnen Datums einer Person, sondern aus der Auswertung einer großen Menge an Daten mit dem Ziel, auf das Verhalten oder den Charakter eines einzelnen Menschen zu schließen.[265]
Diese Leerstelle versucht die EU-Kommission mit ihrem KI-VO-E zu füllen, indem sie eine technologieneutrale und risikobasierte Regulierung Künstlicher Intelligenz vorschlägt. Wie schon zur DSGVO, so kann auch zum KI-VO-E hier keine umfassende Analyse erfolgen.[266] Vielmehr werden wiederum die wichtigsten Regelungen herausgegriffen, die für die Frage des Geheimnisschutzes von KNN relevant werden könnten.
Der Verordnungsentwurf[267] sieht gemäß Art. 1 lit. a Vorschriften vor für „das Inverkehrbringen, die Inbetriebnahme und die Verwendung von Systemen der Künstlichen Intelligenz". Ein KI-System ist in Art. 3 Nr. 1 definiert sind als

> „ein System, das so konzipiert ist, dass es mit Elementen der Autonomie arbeitet, und das auf der Grundlage maschineller und/oder vom Menschen erzeugter Daten und Eingaben durch maschinelles Lernen und/oder logik- und wissensgestützte Konzepte ableitet, wie eine Reihe

[263] Ein Auszug aus der Allgemeinen Ausrichtung des Rates zur KI-Verordnung mit den hier analysierten Normen findet sich im Anhang.

[264] *Dreyer/Schulz*, Was bringt die Datenschutz-Grundverordnung für automatisierte Entscheidungssysteme?, S. 9; vgl. zur Teilhabe insbesondere *Vieth/Wagner*, Teilhabe, ausgerechnet.

[265] *Martini*, Blackbox Algorithmus – Grundfragen einer Regulierung Künstlicher Intelligenz, S. 159.

[266] Siehe für einen Überblick etwa *Ebert/Spiecker gen. Döhmann*, NVwZ 2021, 1188; *Ebers u. a.*, RDi 2021, 528.

[267] Zitiert wird die Allgemeine Ausrichtung vom 6.12.2022, 15698/22.

> von Zielen erreicht wird, und systemgenerierte Ergebnisse wie Inhalte (generative KI-Systeme), Vorhersagen, Empfehlungen oder Entscheidungen hervorbringt, die das Umfeld beeinflussen, mit dem die KI-Systeme interagieren".

Als Methode des maschinellen Lernens sind die hier untersuchten KNN grundsätzlich Gegenstand des Verordnungsentwurfs.
Dies bedeutet jedoch nicht, dass jegliche Anwendung eines KNN einer strengen Regulierung unterworfen würde. Der Verordnungsentwurf sieht vielmehr ein abgestuftes Regulierungssystem vor und unterscheidet vier Risikostufen:
Anwendungen mit „unannehmbare[m] Risiko", die die Werte der Union verletzen, werden in Art. 5 Abs. 1 KI-VO-E verboten.[268] Darunter fallen KI-Systeme zur Verhaltensmanipulation (lit. a und b), zum Social Scoring (lit. c), und, mit Ausnahmen, zur biometrischen Fernidentifikation in Echtzeit durch Strafverfolgungsbehörden (lit. d).
Stark reguliert werden Hochrisiko-KI-Systeme im Sinne des Art. 6 KI-VO-E. Darunter fallen Systeme, die bereits einer der in Anhang II genannten Unionsvorschriften unterfallen und danach auch einer Konformitätsbewertung durch Dritte unterzogen werden müssen, sowie regelmäßig KI-Systeme aus den in Anhang III genannten risikoträchtigen Anwendungsbereichen. Dort finden sich die zu Beginn dieses Kapitels erwähnten kritischen Bereiche wieder, etwa Bildung, Beschäftigung, Kreditwürdigkeitsprüfung, Strafverfolgung und Rechtspflege.
Darüber hinaus sieht der Verordnungsentwurf Regelungen für Anwendungen mit geringem oder minimalem Risiko vor (Transparenzpflichten in Art. 52 und freiwillige Verhaltenskodizes in Art. 69).
Für die Frage des Geheimnisschutzes ist zunächst der Adressatenkreis der Verordnung relevant, aus dem sich mögliche Empfänger geheimer Information ergeben.
Die wichtigsten Adressaten sind hier der Anbieter und der Nutzer. Ersterer ist in Art. 3 Nr. 2 KI-VO-E definiert als:

> „eine natürliche oder juristische Person, Behörde, Einrichtung oder sonstige Stelle, die ein KI-System entwickelt oder entwickeln lässt und dieses System unter dem eigenen Namen oder der eigenen Marke in Verkehr bringt oder in Betrieb nimmt, sei es entgeltlich oder unentgeltlich".

[268] *Europäische Kommission*, Entwurf der KI-Verordnung, COM(2021) 206 final, Begründung, S. 15.

Letzterer bezeichnet gemäß Art. 3 Nr. 4 KI-VO-E

> „eine natürliche oder juristische Person, einschließlich Behörden, Einrichtungen oder sonstige Stellen, unter deren Verantwortung das System verwendet wird".

Aus Sicht des Geheimnisschutzes sind mithin zunächst Transparenzpflichten des Anbieters gegenüber dem Nutzer von Interesse. Hier besteht ein bedeutsamer Unterschied zu den in der DSGVO normierten Informations- und Auskunftsrechten: während diese Rechte im Verhältnis zwischen dem Verantwortlichen und der betroffenen Person greifen, sieht der KI-VO-E keine entsprechenden Rechte des von der KI betroffenen Individuums vor.[269] Zur Durchsetzung von Betroffenenrechten regelt allerdings der Vorschlag für eine Richtlinie über KI-Haftung (KI-Haftungs-RL-E)[270] in Anknüpfung an den KI-VO-E unter anderem die Anordnung zur Offenlegung von Beweismitteln durch nationale Gerichte (Art. 3 KI-Haftungs-RL-E).
Im KI-VO-E selbst regelt lediglich Art. 52 für einen engen Anwendungsbereich Informationspflichten gegenüber natürlichen Personen. Relevant für die Frage des Schutzes von Geschäftsgeheimnissen sind darüber hinaus die Transparenzpflichten des Anbieters gegenüber Aufsichtsbehörden.
Hier soll mit der Darstellung der Transparenzpflichten gemäß Art. 52 KI-VO-E gegenüber natürlichen Personen sowie gemäß Art. 13 KI-VO-E gegenüber Nutzern begonnen werden (private Informationsempfänger), um anschließend solche gegenüber Aufsichtsbehörden und von ihnen beauftragten Dritten zu erörtern (öffentliche Informationsempfänger).

1. Private Informationsempfänger

a) Transparenz gegenüber betroffenen Personen (Artikel 52 KI-VO-E)

Art. 52 KI-VO-E richtet sich an Anbieter (Abs. 1) und Nutzer (Abs. 2, 2a und 3) von KI-Systemen und sieht Transparenzpflichten gegenüber natürlichen Personen vor. Die Norm gilt für alle KI-Systeme, unabhängig von ihrer Risiko-

[269] Siehe dazu kritisch *Ebert/Spiecker gen. Döhmann*, NVwZ 2021, 1188 (1193); *Ebers u. a.*, RDi 2021, 528 (537).

[270] Vorschlag für eine Richtlinie des Europäischen Parlaments und des Rates zur Anpassung der Vorschriften über außervertragliche zivilrechtliche Haftung an künstliche Intelligenz (Richtlinie über KI-Haftung), COM(2022) 496 final.

stufe. Da es sich häufig um eine Verarbeitung personenbezogener Daten handeln wird, dürften die Informationspflichten oft neben den Pflichten aus Art. 13 und 14 DSGVO bestehen.[271]

Gemäß Art. 52 Abs. 1 KI-VO-E muss der Anbieter von KI-Systemen, die für die Interaktionen mit natürlichen Personen bestimmt sind, sicherstellen, dass diese so konzipiert und entwickelt sind, dass natürlichen Personen mitgeteilt wird, dass sie es mit einem KI-System zu tun haben (Ausnahmen gelten für Strafverfolgungsbehörden). Es handelt sich mithin um eine sehr eingeschränkte Transparenzpflicht über das „Ob" des Einsatzes eines KI-Systems, die aus Geheimnisschutzgesichtspunkten nicht relevant sein dürfte.

Gleiches gilt für Art. 52 Abs. 3 KI-VO-E, der in Bezug auf sogenannte *Deepfakes*[272] eine Offenlegung durch den Nutzer dahingehend vorsieht, „dass die Inhalte künstlich erzeugt oder manipuliert wurden." Auch hier ist mithin nur eine Information über das „Ob" des Einsatzes von KI vorgesehen (Ausnahmen gelten unter anderem für Strafverfolgungsbehörden und für künstlerische Zwecke), weshalb Geschäftsgeheimnisse an dem verwendeten Modell ebenfalls nicht berührt sein dürften.

Etwas anderes könnte für die Transparenzpflichten des Art. 52 Abs. 2 und Abs. 2a KI-VO-E gelten, wonach Nutzer eines Systems zur biometrischen Kategorisierung (Abs. 2) oder eines Emotionserkennungssystems (Abs. 2a) betroffene natürliche Personen „über den Betrieb des Systems" informieren muss (Ausnahmen gelten auch hier für Strafverfolgungsbehörden). Die Abweichung im Wortlaut von den Absätzen 1 und 3 könnte dafür sprechen, dass hier nicht nur über das „Ob", sondern auch über das „Wie" („den Betrieb") des Einsatzes von KI informiert werden müsste. Der Wortlaut der Norm selbst ist für eine Klärung unergiebig, könnte aber mangels Spezifizierung eher für eine enge Transparenzpflicht lediglich über das „Ob" des Betriebs sprechen.

Für eine weitere Transparenzpflicht als die in den Absätzen 1 und 3 getroffene könnte allerdings streiten, dass Emotionserkennung und biometrische Kategorisierung von den vier in Art. 52 geregelten Bereichen die eingriffsintensivsten darstellen dürften.

---

[271] *Ebert/Spiecker gen. Döhmann*, NVwZ 2021, 1188 (1191); siehe zum Verhältnis des KI-VO-E zu den Informations- und Auskunftsrechten der DSGVO: Entwurf der KI-Verordnung, Allgemeine Ausrichtung, Erwägungsgrund 58a.

[272] In Art. 52 Abs. 3 KI-VO-E definiert als „Bild-, Ton- oder Videoinhalte [...], die wirklichen Personen, Gegenständen, Orten oder anderen Einrichtungen oder Ereignissen merklich ähneln und einer Person fälschlicherweise als echt oder wahrhaftig erscheinen würden".

Auch systematisch könnte der Vergleich mit Art. 13 Abs. 1 KI-VO-E eine weitergehende Regelung nahelegen, denn dort wird vorgeschrieben, dass der „Betrieb" eines KI-Systems „hinreichend transparent" sein muss, damit die Nutzer das System angemessen verstehen und verwenden können.
Auch Art. 29 Abs. 4 KI-VO-E, nach dem die Nutzer eines Hochrisiko-KI-Systems dessen Betrieb anhand der Gebrauchsanweisung überwachen sollen, geht in diese Richtung. Beide Normen verwenden den Begriff „Betrieb" in Zusammenhang mit einem Verständnis über die Funktionsweise.
Allerdings enthält Art. 52 Abs. 2 und Abs. 2a KI-VO-E gerade keinen entsprechenden Zusatz zum Zweck der Transparenz oder einen anderen Hinweis über die Reichweite der Informationspflicht. Zudem spricht die Begründung der Norm gegen eine Informationspflicht über das „Ob" des Einsatzes hinaus. Dort heißt es: „Interagieren Personen mit KI-Systemen oder werden deren Emotionen oder Merkmale durch automatisierte Mittel erkannt, müssen die Menschen hierüber informiert werden."[273] „Hierüber" bezieht sich recht eindeutig lediglich auf den Umstand, dass Emotionen durch automatisierte Mittel erkannt werden.
Im Ergebnis wird hier daher davon ausgegangen, dass auch die Absätze 2 und 2a nur eine Transparenzpflicht hinsichtlich des Einsatzes an sich vorsieht. Es besteht jedoch angesichts der dargestellten Unsicherheiten in der Auslegung keine ausreichende Rechtssicherheit.
Zusammenfassend lässt sich daher festhalten, dass auch die in Art. 52 KI-VO-E vorgesehenen Transparenzpflichten gegenüber natürlichen Personen für Geschäftsgeheimnisse am betreffenden System ungefährlich sein dürften.[274]

b) Transparenz gegenüber Nutzern (Artikel 13 KI-VO-E)

Anders verhält es sich mit den Transparenzpflichten des Anbieters gegenüber dem Nutzer eines KI-Systems.
Hier ist Art. 13 KI-VO-E einschlägig, der für Hochrisiko-KI-Systeme gilt. Nach Absatz 1 müssen diese so konzipiert und entwickelt werden, „dass ihr Betrieb hinreichend transparent ist, damit die Nutzer und Anbieter ihre in Kapitel 3

[273] *Europäische Kommission*, Entwurf der KI-Verordnung, COM(2021) 206 final, Begründung S. 17.

[274] Lediglich ein mögliches Geschäftsgeheimnis an der Information, dass ein entsprechendes KI-System genutzt wird, würde durch die Transparenzverpflichtungen offenbar. Siehe zur Information über den Einsatz von KI als Geschäftsgeheimnis auch *Alexander*, MMR 2021, 690 (691).

dieses Titels festgelegten einschlägigen Pflichten erfüllen können und damit die Nutzer das System angemessen verstehen und verwenden können."
Der Wortlaut der Norm verdeutlich, dass hier Transparenzpflichten hinsichtlich des „Wie" des Einsatzes von KI getroffen werden, da sie auf unter anderem auf das Verständnis und die Verwendung des Systems zielen. Allerdings bleiben die Anforderungen an den Anbieter auch hier denkbar vage. Der Gebrauch einer Vielzahl unbestimmter Rechtsbegriffe („hinreichend transparent", „angemessen verstehen und verwenden") lässt den Anbieter über seine aus der Norm erwachsenden Pflichten im Dunkeln und dürfte sich als nicht praxistauglich herausstellen.[275]
Die Pflicht zur Bereitstellung einer Gebrauchsanweisung gemäß Art. 13 Abs. 2 KI-VO-E wird durch Absatz 3 schon konkreter, der eine Liste mit Informationen enthält, die dem Nutzer in der Gebrauchsanweisung zur Verfügung gestellt werden müssen. Für die vorliegende Fragestellung dürfte insbesondere die Information nach Art. 13 Abs. 2 lit. b KI-VO-E interessant sein, die unter anderem „die Merkmale, Fähigkeiten und Leistungsgrenzen des Hochrisiko-KI-Systems" enthalten soll. Was jedoch insbesondere unter die „Merkmale" eines KI-Systems fällt, bleibt offen. Die unter i) bis v) genannten Informationen dürften eine nicht abschließende Liste solcher „Merkmale, Fähigkeiten und Leistungsgrenzen" darstellen („einschließlich"). Keine der dort genannten Informationen dürfte jedoch für die Frage des Geheimnisschutzes an einem trainierten KI-Modell, genauer seine Reproduzierbarkeit, relevant sein.
Der Regelanwender wird mithin auf den Begriff des „Merkmals" in lit. b zurückgeworfen und muss sich fragen, welche Information über sein KI-System darunter zu subsumieren ist.
Zunächst kann festgehalten werden, dass Gebrauchsanweisungen gemäß Art. 13 Abs. 2 KI-VO-E „präzise, vollständige, korrekte und eindeutige Informationen in einer für die Nutzer relevanten [...] und verständlichen Form enthalten" müssen.
Systematisch stehen das in derselben Norm geregelte Transparenzgebot und der Inhalt der Gebrauchsanweisung in einem Zweckverhältnis zueinander, sodass also die Gebrauchsanweisung auch der Transparenz dienen soll.[276]

[275] Siehe zu diesem Defizit auch *Ebers u. a.*, RDi 2021, 528 (533 f.); *Roos/Weitz*, MMR 2021, 844 (847), die sich allerdings noch auf den Kommissions-Entwurf vom 21.04.2021 beziehen.
[276] Vgl. Entwurf der KI-Verordnung, Allgemeine Ausrichtung, Erwägungsgrund 47.

Zu beachten ist außerdem, dass Art. 13 KI-VO-E, anders als die Informations- und Auskunftsrechte der Art. 13-15 DSGVO, nicht die betroffene Person adressiert, sondern den Nutzer des KI-Systems. Der Empfängerhorizont dürfte mithin ein anderer sein als in der DSGVO und die Information nicht zwingend auf ein für Laien verständliches Maß begrenzt sein. Denn die Verordnung soll nicht gelten für die Pflichten von Nutzern, die natürliche Personen sind und KI-Systeme im Rahmen einer ausschließlich persönlichen und nicht beruflichen Tätigkeit verwenden gemäß Art. 2 Nr. 8 KI-VO-E.
Zwar lässt sich daraus nicht eindeutig schließen, dass die Gebrauchsanweisung nur Informationen für Nutzer enthalten muss, die das KI-System im Rahmen ihrer beruflichen Tätigkeit nutzen.[277] Denn die Gebrauchsanweisung kann auch schlicht zur Information privater Nutzer dienen, ohne dass diese Pflichten aus dem KI-VO-E unterlägen.[278] Sie muss jedoch jedenfalls auch Informationen für Nutzer enthalten, die das KI-System beruflich nutzen. Es ist anzunehmen, dass diese dann auch über fachlich kompetentes Personal verfügen, das auch komplexere technische Erklärungen der „Merkmale" eines KI-Systems zur Interpretation seiner Ergebnisse verwenden könnte. Dafür spricht auch, dass die in Art. 14 KI-VO-E vorgesehene menschliche Aufsicht von Hochrisiko-KI-Systemen auch durch Nutzer vorgenommen werden kann (vgl. Art. 14 Abs. 3 lit. b)) und dass die aufsichtspflichtige Person „die Fähigkeiten und Grenzen des Hochrisiko-KI-Systems vollständig verstehen" können soll (vgl. Art. 14 Abs. 4 lit. a)).[279]
In Anbetracht der vorstehenden Überlegungen ist es nicht ausgeschlossen, dass „Merkmale" des KI-Systems im Sinne des Art. 13 Abs. 3 lit. a KI-VO-E auch Informationen eines KI-Modells umfassen könnten, die als Geschäftsgeheimnis geschützt sind. In welcher Form diese im Falle eines KNN dargestellt werden könnten und im Hinblick auf den Normzweck werden sollten wird in den letzten beiden Teilen dieser Arbeit analysiert.

---

[277] So im ursprünglichen Entwurf vorgesehen, wo der Nutzer in Art. 3 Nr. 4 bereits tatbestandlich und unabhängig von seinen Pflichten entsprechend eingeschränkt war, vgl. *Europäische Kommission*, Entwurf der KI-Verordnung, COM(2021) 206 final, Art. 3 Nr. 4.

[278] Diese Unschärfe hinsichtlich des Empfängerkreises ist mit Blick auf eine rechtssichere Auslegung des Umfangs der in der Gebrauchsanweisung zur Verfügung zu stellenden Informationen äußerst problematisch.

[279] Die Frage, ob sich aus Art. 14 KI-VO-E nicht strenggenommen ein Verbot von Blackbox-Systemen ergeben müsste, werfen Ebers et. al. auf: *Ebers u. a.*, RDi 2021, 528 (534).

2. Öffentliche Informationsempfänger

Die dritte hier zu untersuchende Konstellation betrifft Transparenzerfordernisse gegenüber Behörden und von ihnen beauftragten Dritten.

a) Überblick

Hier können grundsätzlich zwei Gruppen von Empfängern der von Transparenzpflichten betroffenen Informationen unterschieden werden: die notifizierten Stellen und die zuständigen nationalen Behörden.
Notifizierte Stellen sind Konformitätsbewertungsstellen, die gemäß dem KI-VO-E und anderen einschlägigen Harmonisierungsvorschriften der Union benannt wurden (vgl. Art. 3 Nr. 22 KI-VO-E).
Die notifizierten Stellen werden von den notifizierenden Behörden (Art. 3 Nr. 20, Art. 30 KI-VO-E) auf Antrag (Art. 31 KI-VO-E) durch ein Notifizierungsverfahren (Art. 32 KI-VO-E) anerkannt. Sie sind gemäß Art. 34a KI-VO-E für die Überprüfung der Konformität von KI-Hochrisiko-Systemen nach den in Art. 43 KI-VO-E genannten Konformitätsbewertungsverfahren zuständig. Sie müssen nach Art. 33 KI-VO-E eine Vielzahl von Kriterien erfüllen und insbesondere Maßnahmen zur Wahrung der Vertraulichkeit treffen gemäß Art. 33 Abs. 6 S. 1 KI-VO-E. Darüber hinaus unterliegen Informationen, von denen Mitarbeiter der notifizierten Stellen bei der Durchführung ihrer Aufgaben gemäß dem KI-VO-E Kenntnis erlangen, der beruflichen Schweigepflicht (Art. 33 Abs. 6 S. 2 KI-VO-E). Die Verpflichtung zur Wahrung der Vertraulichkeit, auch explizit in Bezug auf Geschäftsgeheimnisse, ist zudem in Art. 70 Abs. 1 lit. a KI-VO-E geregelt.[280] Auch die notifizierenden Behörden, denen die notifizierten Stellen alle einschlägigen Unterlagen – auch die des Anbieters – zugänglich machen müssen (Art. 34a Abs. 3 KI-VO-E), müssen ihrerseits die Vertraulichkeit der erlangten Information gewährleisten (Art. 30 Abs. 6 i. V. m. Art. 70 KI-VO-E).
Dennoch birgt die Auslagerung der Konformitätsbewertung an die notifizierten Stellen und mithin an Dritte ein höheres Risiko der Offenbarung von Geschäftsgeheimnissen.[281] Dieses Risiko wird noch vergrößert durch die den noti-

[280] Zu Art. 70 KI-VO-E näher unten.

[281] So auch bereits vor Veröffentlichung des KI-VO-E *Martini*, Blackbox Algorithmus – Grundfragen einer Regulierung Künstlicher Intelligenz, S. 229; siehe zum Risiko der Offenbarung von Geschäftsgeheimnissen gegenüber Prüfern *Huber*, in: Ann/Loschelder/Grosch, Praxishandbuch Know-how-Schutz, S. 594.

fizierten Stellen eingeräumte Möglichkeit, ihrerseits bestimmte Aufgaben an Zweigstellen zu übertragen oder an Unterauftragsnehmer zu vergeben (Art. 34 KI-VO-E). Diese unterliegen dann zwar ebenfalls gemäß Art. 34 Abs. 1 KI-VO-E den Anforderungen des Art. 33 KI-VO-E, dennoch wird der Adressatenkreis potenzielle geheimer Information durch diese Möglichkeit stark ausgeweitet und unübersichtlicher.
Weiterer Empfänger von geheimer Information sind die zuständigen nationalen Behörden. Damit sind gemeint die notifizierende Behörde und die Marktüberwachungsbehörde (vgl. Art. 3 Nr. 43 KI-VO-E). Sie sind ebenfalls gemäß Art. 70 Abs. 1 KI-VO-E zur Wahrung der Vertraulichkeit der erlangten Information verpflichtet.
Die wesentlichen Transparenzverpflichtungen gegenüber Behörden und behördlich berufenen Stellen erwachsen aus dem Konformitätsbewertungsverfahren gemäß Art. 43 KI-VO-E und den Zugangsrechten für Marktüberwachungsbehörden und nationale Behörden oder öffentlichen Stellen zur Einhaltung des Unionsrechts zum Schutz der Grundrechte gemäß den Artikel 63 und 64 KI-VO-E.

b) Konformitätsbewertung und technische Dokumentation (Artikel 43 und 11 KI-VO-E)

Der KI-VO-E sieht vor, dass Konformitätsbewertungen von Hochrisiko-KI-Systemen grundsätzlich intern durch den Anbieter „in eigener Verantwortung" durchgeführt werden (vgl. Erwägungsgrund 64, Art. 16 lit. e, 19, 43 Abs. 2 i. V. m. Anhang VI).[282] Eine Ausnahme gilt für Hochrisiko-KI-Systeme zur biometrischen Fernidentifizierung von Personen, die einem (externen) Konformitätsbewertungsverfahren durch eine notifizierte Stelle unterliegen (vgl. Erwägungsgrund 64, Art. 16 lit. e, 19, 43 Abs. 1 i. V. m. Anhang VII).
Im Falle der internen Konformitätsbewertung bestehen naturgemäß grundsätzlich keine Risiken für den Geheimnisschutz. Auch für die Anbieter der betreffenden Hochrisiko-KI-Systeme bestehen jedoch Kooperationspflichten, die

[282] Die Norm ist auf viel Kritik gestoßen, vgl. nur: *Ebert/Spiecker gen. Döhmann*, NVwZ 2021, 1188 (1193); *Ebers u. a.*, RDi 2021, 528 (537); dagegen die Norm begrüßend *Roos/Weitz*, MMR 2021, 844 (848); siehe zur Forderung umfassender Zulassungskontrollen für besonders risikoträchtige Anwendungen Künstlicher Intelligenz bereits vor Veröffentlichung des KI-VO-E *Martini*, Blackbox Algorithmus – Grundfragen einer Regulierung Künstlicher Intelligenz, S. 225 ff.

sehr wohl auf eine Offenbarung von Geschäftsgeheimnissen herauslaufen können. So verpflichtet Art. 23 Satz 1 KI-VO-E den Anbieter eines Hochrisiko-KI-Systems, einer zuständigen nationalen Behörde die für die Bewertung der Konformität des Systems mit Kapitel 2 erforderlichen Informationen und Unterlagen zu übermitteln. Die „Informationen und Unterlagen" dürften die technische Dokumentation gemäß Art. 11 KI-VO-E einschließen.
Die technische Dokumentation bildet eine der beiden Grundlagen des externen Konformitätsbewertungsverfahren (Anhang VII Nr. 1). Ihr Mindestinhalt ist in Anhang IV des KI-VO-E geregelt (vgl. Art. 11 Abs. 1 Satz 3 KI-VO-E), dessen detaillierte Vorgaben für die Frage der Offenbarung von Geschäftsgeheimnissen von Interesse sind.
Nach Nr. 2 des Anhangs IV muss die technische Dokumentation unter anderem eine „[d]etaillierte Beschreibung der Bestandteile des KI-Systems" enthalten. Dazu zählen gemäß lit. b unter anderem:

> „Entwurfsspezifikationen des Systems, insbesondere die allgemeine Logik des KI-Systems und der Algorithmen; wichtigste Entwurfsentscheidungen mit den Gründen und Annahmen, auch in Bezug auf Personen oder Personengruppen, auf die das System angewandt werden soll; hauptsächliche Klassifizierungsentscheidungen; was das System optimieren soll und welche Bedeutung den verschiedenen Parametern dabei zukommt",

und gemäß lit. c zudem insbesondere die

> „Beschreibung der Systemarchitektur, aus der hervorgeht, wie Softwarekomponenten aufeinander aufbauen oder einander zuarbeiten und in die Gesamtverarbeitung integriert sind".

Darüber hinaus werden in lit. a und lit. d genaue Angaben zu Entwicklung, Trainingsdatensätzen und Training des Systems gefordert.
Aus den genannten Beispielen lässt sich schließen, dass die technische Dokumentation eine umfassende Beschreibung des KI-Systems enthalten muss. Da neben der „allgemeinen Logik" des Systems und der „Algorithmen" auch, „hauptsächliche Klassifizierungsentscheidungen", Angaben zu Parametern und deren Bedeutung für die Optimierungsentscheidung des Systems erfasst sind, dürfte für Künstliche Neuronale Netze die Offenlegung der Netzarchitektur und möglicherweise auch der Gewichtungen erforderlich sein. Denn die Bedeu-

tung der Parameter für die Optimierungsentscheidung des Systems wird sich ohne Kenntnis der Gewichte schwerlich überprüfen lassen.
Dafür spricht auch, dass das in Anhang VII beschriebene Verfahren der im Rahmen der externen Konformitätsbewertung vorzunehmenden Kontrolle der technischen Dokumentation auf eine umfassende Prüfung des KI-Systems abzielt. Sogar Zugang zum Quellcode soll den notifizierten Stellen auf begründeten Antrag gewährt werden, sofern dies zur Konformitätsbewertung notwendig ist und Tests und Überprüfungen aufgrund der vom Anbieter bereitgestellten Daten und Dokumentation ausgeschöpft wurden oder sich als unzureichend erwiesen haben (Nummer 4.5.). Auch die in den Nummern 4.3. und 4.4. des Anhangs VII niedergelegten Rechte auf Zugang zu Trainings- und Testdaten und auf Durchführung von Tests des KI-Systems verdeutlichen den klaren Willen des EU-Gesetzgebers, die notifizierten Stellen umfassend über das jeweilige KI-System zu informieren. Dass der Quellcode eines KNN die wesentliche Information des Netzes – dessen Gewichte – gerade nicht enthält,[283] dürfte nicht gegen ein Dokumentationserfordernis auch der Gewichte sprechen. Vielmehr ist davon auszugehen, dass der Zugriff auf den Quellcode, wie bei klassischer Software und GOFAI der Fall, als maximale Offenlegung des Systems gemeint ist.
Im Rahmen der Konformitätsbewertung sollen mithin umfassende Informationen über das jeweilige KI-System offengelegt werden, bei denen es sich in vielen Fällen um Geschäftsgeheimnisse handeln wird. Adressaten dieser in der technischen Dokumentation enthaltenen Informationen sind die zuständigen nationalen Behörden und die notifizierten Stellen (vgl. Art. 11 Abs. 1 Satz 2 KI-VO-E).
In welchem Maße die Erfüllung der beschriebenen Anforderungen speziell für Künstliche Neuronale Netze mit der Offenlegung von Geschäftsgeheimnissen verbunden ist, wird im dritten Teil dieser Arbeit untersucht.

c) Kontrolle durch Behörden (Artikel 63 und 64 KI-VO-E)

Weitere Transparenzpflichten gegenüber Behörden sehen die Art. 63 und 64 KI-VO-E vor.
Art. 63 KI-VO-E enthält Befugnisse der Marktüberwachungsbehörde. Dabei handelt es sich um die „nationale Behörde, die die Tätigkeit durchführt und die Maßnahmen ergreift, die in der Verordnung (EU) 2019/1020 vorgesehen sind“

[283] Siehe dazu unten, Kapitel 5 A.

(Art. 3 Nr. 26 KI-VO-E).[284] Ihr wird in Absatz 8 umfangreicher Zugang zu Dokumentation und Datensätzen gewährt, sofern dies relevant und zur Wahrnehmung ihrer Aufgaben erforderlich ist. Darüber hinausgehend kann die Marktüberwachungsbehörde gemäß Art. 63 Abs. 9 KI-VO-E unter zwei Voraussetzung Zugang zum Quellcode verlangen: der Zugang muss zur Bewertung der Konformität eines Hochrisiko-KI-System mit den Anforderungen des KI-VO-E erforderlich sein und Tests und Überprüfungen aufgrund der vom Anbieter bereitgestellten Daten und Dokumentation wurden ausgeschöpft oder haben sich als unzureichend erwiesen. Art. 63 Abs. 10 KI-VO-E stellt klar, dass die Marktüberwachungsbehörde den Vertraulichkeitspflichten des Art. 70 KI-VO-E unterliegt.

Art. 64 KI-VO-E räumt noch weiteren Behörden Zugangsrechte ein: gemäß Absatz 3 können nationale Behörden oder öffentliche Stellen, die die Anwendung des Unionsrechts zum Schutz der Grundrechte in Bezug auf den Einsatz der in Anhang III aufgeführten Hochrisiko-KI-Systeme überwachen oder durchsetzen, alle auf der Grundlage der Verordnung erstellten oder geführten Unterlagen anfordern und einsehen, sofern dies für die Ausübung ihres Auftrags im Rahmen ihrer Befugnisse notwendig ist. Zu diesen nationalen Behörden und Stellen sind gemäß Erwägungsgrund 79a auch die Gleichstellungsstellen und Datenschutzbehörden zu zählen. Sie können bei der Marktüberwachungsbehörde auch die Durchführung von Tests von KI-Systemen beantragen gemäß Art. 64 Abs. 5 KI-VO-E.

Artikel 63 KI-VO-E gewährt mithin zunächst den Marktüberwachungsbehörden umfangreiche Zugangsrechte, insbesondere zum Quellcode eines KI-Systems. Artikel 64 erweitert jedoch den Empfängerkreis potenziell geheimer Information auf weitere Behörden und öffentliche Stellen. Auch wenn diese nicht Zugang zum Quellcode bekommen sollen, so dürften die in Art. 64 Abs. 3 KI-VO-E genannten „Unterlagen" auch die technische Information und mithin alle relevanten Informationen über ein KI-System umfassen. Natürlich unterliegen auch die mit Grundrechtsschutz befassten Behörden und öffentlichen Stellen den Vertraulichkeitspflichten des Art. 70 KI-VO-E (so explizit Art. 64 Abs. 6 KI-VO-E). Dennoch birgt jede Erweiterung des Empfängerkreises möglicher-

[284] Es handelt sich um die Verordnung über Marktüberwachung und die Konformität von Produkten sowie zur Änderung der Richtlinie 2004/42/EG und der Verordnungen (EG) Nr. 765/2008 und (EU) Nr. 305/2011.

weise geheimer Information das Risiko einer Offenlegung von Geschäftsgeheimnissen.

3. Vertraulichkeit (Artikel 70 KI-VO-E)

Die Analyse des KI-VO-E im Hinblick auf Transparenzpflichten gegenüber Privaten und gegenüber Behörden und behördlich beauftragten Dritten hat gezeigt, dass eine Offenlegung von Geschäftsgeheimnissen gegenüber der ersten Gruppe zumindest nicht ausgeschlossen ist und dass sie eine zwingende Folge der Zugangsrechte der zweiten Gruppe ist.
Anders als die DSGVO, die den Schutz von geistigem Eigentum und von Geschäftsgeheimnissen lediglich in den Erwägungsgründen adressiert, sieht der KI-VO-E mit Art. 70 eine Regelung über Vertraulichkeit im Umgang mit Information vor:

> „Die zuständigen nationalen Behörden, die notifizierten Stellen, die Kommission, der KI-Ausschuss und alle anderen natürlichen oder juristischen Personen, die an der Anwendung dieser Verordnung beteiligt sind, ergreifen im Einklang mit dem Unionsrecht oder dem nationalen Recht geeignete technische und organisatorische Maßnahmen, um die Vertraulichkeit der Informationen und Daten, in deren Besitz sie bei der Ausführung ihrer Aufgaben und Tätigkeiten gelangen, sicherzustellen" (Art. 70 Abs. 1 KI-VO-E).

Es folgt eine Aufzählung der konkret durch die Vertraulichkeit zu schützenden Interessen, an erster Stelle "Rechte des geistigen Eigentums, vertrauliche Geschäftsinformationen oder Geschäftsgeheimnisse natürlicher oder juristischer Personen, auch Quellcodes, [...]". Die Nennung speziell des Quellcodes zeigt, dass sich der Gesetzgeber der Europäischen Union der Kritikalität dieser Information für die Unternehmen bewusst ist. Gleichzeitig gilt speziell für KNN, dass der Quellcode die Trainingsergebnisse des Netzes, und damit einen wichtigen Teil seiner Funktionalität, gerade nicht enthält.[285]

[285] Dazu näher unten.

Ob die Vertraulichkeitspflichten von öffentlichen und notifizierten Stellen ein angemessenes Gegengewicht zu den umfangreichen Transparenzverpflichtungen des Verordnungsentwurfs bilden, darf zumindest bezweifelt werden.[286]
Als Fortschritt gegenüber dem ursprünglichen Kommissions-Entwurf ist es jedenfalls zu werten, dass Art. 70 KI-VO-E über zuständige nationale Behörden und notifizierte Stellen hinaus nun alle an der Anwendung der Verordnung beteiligten natürlichen und juristischen Personen in den Kreis der Verpflichteten aufnimmt. Ob dies auch die Nutzer einschließt, ist zwar zweifelhaft, denn die „Anwendung der Verordnung" dürfte wohl in erster Linie in Händen der Aufsichts- und Durchsetzungsbehörden liegen. Allerdings wäre eine Anwendung der Norm auf Nutzer zu begrüßen, da diese für den Anbieter eines KI-Systems die kritischste Empfängergruppe von geheimer Information darstellen.[287] Eine entsprechende Regelung, die Artikel 70 für den Informationsaustausch zwischen Anbietern für anwendbar erklärt, findet sich in Art. 4b Abs. 5 KI-VO-E.
Im Begründungsteil des Verordnungsentwurfs präzisiert die Kommission ihre Vorstellung des Geheimnisschutzes unter der KI-Verordnung:

> „Auch die Pflicht zu größerer Transparenz wird das Recht auf Schutz des geistigen Eigentums (Artikel 17 Absatz 2) nicht unverhältnismäßig beeinträchtigen, da sie auf die Mindestinformationen beschränkt ist, die eine Person benötigt, um ihr Recht auf einen wirksamen Rechtsbehelf auszuüben, sowie auf die Transparenz, die Aufsichts- und Durchsetzungsbehörden im Rahmen ihrer Aufgaben benötigen. Jede Offenlegung von Informationen erfolgt unter Einhaltung der einschlägigen Rechtsvorschriften, auch der Richtlinie (EU) 2016/943 über den Schutz vertraulichen Know-hows und vertraulicher Geschäftsinformationen (Geschäftsgeheimnisse) vor rechtswidrigem Erwerb sowie rechtswidriger Nutzung und Offenlegung. Benötigen öffentliche und notifizierte Stellen für die Prüfung der Einhaltung der wesentlichen Pflichten Zugang zu vertraulichen Informationen oder zu Quellcodes, sind sie zur Wahrung der Vertraulichkeit verpflichtet."[288]

---

[286] So bezüglich der Offenlegung des Quellcodes gegenüber Marktüberwachungsbehörden *Ebers u. a.*, RDi 2021, 528 (535).

[287] Siehe zur Gruppe der Wettbewerber als Risiko für Abfluss von Know-how *Huber*, in: Ann/Loschelder/Grosch, Praxishandbuch Know-how-Schutz, S. 592 f.

[288] *Europäische Kommission*, Entwurf der KI-Verordnung, COM(2021) 206 final, S. 13.

Der erste Teil des Absatzes lässt eine Abwägungsentscheidung vermuten zwischen Transparenz und Schutz des geistigen Eigentums und von Geschäftsgeheimnissen. Allerdings scheint im Zweifel der Vorzug der Transparenz gegeben zu werden, da zumindest eine nicht unverhältnismäßige Beeinträchtigung des geistigen Eigentums hingenommen wird. Wie sich der Konflikt zwischen Transparenz und Geheimnisschutz auflösen soll, bleibt hingegen gänzlich unbeantwortet: es wird schlicht postuliert, dass die Offenlegung von Informationen unter Einhaltung der GeschGeh-RL erfolgt, was aufgrund des rein faktischen Schutzes von Geschäftsgeheimnissen durch die Richtlinie nahezu unmöglich erscheint.
Auch Erwägungsgrund 83 des KI-VO-E, nach dem

> „alle an der Anwendung dieser Verordnung beteiligten Parteien im Einklang mit dem Unionsrecht und dem nationalen Recht die Vertraulichkeit der im Rahmen der Durchführung ihrer Tätigkeiten erlangten Informationen und Daten wahren“

sollten, kann aus Sicht des Geheimnisschutzes nur als reines Lippenbekenntnis gewertet werden.

4. Ergebnis

Im Ergebnis lässt sich festhalten, dass der KI-VO-E sehr umfangreiche Transparenzpflichten gegenüber Behörden und von ihnen notifizierten Stellen vorsieht. Die Transparenzpflichten gegenüber Privaten fallen geringer aus, sind jedoch in ihrem Umfang unklar.
Zumindest für die Gruppe der öffentlichen Informationsempfänger werden Vertraulichkeitspflichten, insbesondere die Verpflichtung zum Schutz von geistigem Eigentum und Geschäftsgeheimnissen, explizit hervorgehoben. Ob die Gefahr des Verlustes von Geschäftsgeheimnissen dadurch vermieden werden kann, darf bezweifelt werden. Mit Blick auf private Informationsempfänger, die der Vertraulichkeitspflicht wohl nicht unterliegen, ist klar: der faktische Schutz von Geschäftsgeheimnissen ist hier erheblich gefährdet.

D. Fazit

Sowohl die Informations- und Auskunftsrechte der DSGVO als auch die Transparenzverpflichtungen aus dem KI-VO-E können die Offenlegung potenziell als Geschäftsgeheimnis geschützter Information von KI-Systemen erfordern.
Während die Transparenzpflichten des KI-VO-E gegenüber Behörden und notifizierten Stellen recht klare Angaben zu Art und Inhalt der Information enthalten, die zur Verfügung gestellt werden soll, bleiben die entsprechenden Angaben in der DSGVO, aber auch in den Transparenzpflichten des KI-VO-E gegenüber Nutzern, unbestimmt.
Ebenso bleibt über beide Regelungswerke weitestgehend unklar, wie sich Transparenzpflichten und Geheimnisschutz zueinander verhalten und wie aus den widerstreitenden Interessen entstehende Konflikte aufgelöst werden sollen.
Insbesondere §§ 1 Abs. 2 und 3 Abs. 2 GeschGehG i. V. m. Art. 5 lit. d RL (EU) 2016/943 zeigen zwar Möglichkeiten auf, in denen die Erlangung, Nutzung oder Offenlegung von Geschäftsgeheimnissen erlaubt sein kann. Das Verhältnis von Geschäftsgeheimnissen und Informationspflichten bleibt jedoch dennoch unklar.[289]
Da explizite Regelungen zu einem Vorrangverhältnis fehlen, der Schutz von Geschäftsgeheimnissen jedoch durchaus anerkannt wird, werden sich Konflikte nur über eine Interessenabwägung lösen lassen. Diese Interessenabwägung lässt sich auf den Verhältnismäßigkeitsgrundsatz im Unionsrecht gründen und ist bei Transparenzpflichten üblich.[290]
Der Fokus dieser Arbeit liegt jedoch nicht auf dieser Interessenabwägung, sondern vielmehr auf einer notwendigen Vorfrage: um die Interessen der zu Transparenz Verpflichteten angemessen abwägen zu können, müssen diese Interessen zunächst einmal herausgearbeitet werden. Denn erst wenn Klarheit darüber besteht, ob und in welchem Umfang Geschäftsgeheimnisse betroffen sind, können sie in die Abwägung eingestellt werden.[291] Dieser Aspekt wird in der Debatte um Transparenz und Geheimnisschutz bisher vernachlässigt, er ist jedoch essentiell. Die Schwierigkeit besteht darin, dass die Frage nicht technologieneutral beantwortet werden kann und zudem eine Einzelfallbetrachtung erforder-

[289] *Alexander*, MMR 2021, 690 (690).
[290] *Alexander*, MMR 2021, 690 (693 ff.) der eine dreistufige, „strukturierte Interessenabwägung" anmahnt.
[291] So auch *Alexander*, MMR 2021, 690 (695).

lich macht. Dennoch können und müssen für verschiedene Technologien beziehungsweise KI-Systeme abstrakt diejenigen Informationen herausgearbeitet werden, die als Geschäftsgeheimnis geschützt sein können und anhand derer dann im Einzelfall eine Abwägung erfolgen kann. Die nachfolgenden Kapitel widmen sich daher der Frage, was die Information eines Künstlichen Neuronalen Netzes ausmacht, ob sie als Geschäftsgeheimnis geschützt sein kann und wie sie zur Erfüllung von Transparenzpflichten dargestellt werden kann.

Teil 3

Information eines Künstlichen Neuronalen Netzes: Darstellung und Schutz

Im Rahmen von Transparenzpflichten in der deutschen und EU-Gesetzgebung wird grundsätzlich keine vollumfängliche Offenlegung gefordert, sondern die Transparenz steht üblicherweise unter dem Vorbehalt des Schutzes von Geschäftsgeheimnissen. Und auch in der juristischen Forschung zu Künstlichen Neuronalen Netzen wird vor diesem Hintergrund der besondere Stellenwert des Geheimnisschutzes betont. Eine genaue Analyse, was genau das Geschäftsgeheimnis an einem KNN oder andern ML-Modellen ausmacht und wie weit der Schutzumfang ist, wird nicht vorgenommen. Sie ist jedoch erforderlich, da sie gewissermaßen das Spiegelbild der unterschiedlichen Möglichkeiten bildet, wie den Transparenzpflichten nachgekommen werden kann.

Um die implizit im Netz gespeicherte Information explizit zu machen, muss sie symbolisch repräsentiert werden.[292] Dies geschieht anhand von Zeichen, welche die implizite Information repräsentieren. Durch diesen Vorgang kann die Information aus dem System herausgelöst und für den menschlichen Betrachter erfassbar gemacht werden.[293]

Aufbauend auf den im ersten und zweiten Teil gelegten mathematischen und informationstheoretischen Grundlagen wird in diesem Teil herausgearbeitet, welche semantische Information ein KNN enthält und auf welche Weise sie dargestellt werden kann.

Die Betrachtung erfolgt dabei stufenweise von der konkretesten zur abstraktesten Darstellung der Information eines KNN.

Zunächst wird analysiert, was genau die implizit im KNN repräsentierte, semantische Information des KNN ausmacht, die häufig als sein „Wissen“ bezeichnet wird. Daran knüpft auch die Untersuchung der Schutzfähigkeit des KNN als Geschäftsgeheimnis an, denn das GeschGehG schützt semantische Information, unabhängig von der Art ihrer Verkörperung. Anschließend werden weitere immaterialgüterrechtliche Schutzmöglichkeiten der semantischen Information des KNN geprüft.

Danach werden die verschiedenen Darstellungsmöglichkeiten vorgestellt und untersucht, ob die Information des KNN in ihnen explizit wird und sich somit zur Erklärung gegenüber einem Laien eignet. Ebenso wird untersucht, ob die jeweilige Darstellungsform zur Offenlegung des Geschäftsgeheimnisses und zur Reproduzierbarkeit des KNN führen. Dies führt zur Frage des immaterialgüter-

---

[292] Siehe etwa *Zech*, Weizenbaum Series 2020, 1 (13 ff.).

[293] Zech spricht von der „Abstraktion der Information vom System“, *Zech*, Information als Schutzgegenstand, S. 24.

rechtlichen Schutzes der jeweiligen Darstellungsform. Da es sich um syntaktische Information handelt, beschränkt sich diese Prüfung auf den urheberrechtlichen Schutz.
Ziel dieses Teils ist es mithin, den Bereich zu finden, an dem sich Darstellungsformen mit der Offenbarung von Geschäftsgeheimnissen überschneiden. Hier spielt die Dichotomie zwischen Erklärbarkeit und Reproduzierbarkeit die entscheidende Rolle: nicht alle Information, welche die Funktionsweise eines KNN – sei es abstrakt oder anhand einer konkreten Entscheidung – erklärt, führt zur Offenbarung des Geschäftsgeheimnisses und ist ausreichend, um ein Modell nachzubauen.

Kapitel 4

Semantische Information eines Künstlichen Neuronalen Netzes

A. Untersuchungsgegenstand

Nach der *Physical Symbol System Hypothesis* ist nicht nur die Information in einem System durch Symbole repräsentiert, sondern die Regeln und Prozesse der Informationsverarbeitung selbst können ihrerseits symbolisch repräsentiert werden.[294] So kann in einem klassischen Computerprogramm die Informationsverarbeitung am Algorithmus abgelesen werden, ist also durch den Programmierer vorgegeben.[295]
Anders in einem Künstlichen Neuronalen Netz: die Informationsverarbeitung erfolgt dort nicht anhand eines Algorithmus. Zwar sind Algorithmen Teil eines KNN – die Regeln der Informationsverarbeitung selbst erlernt das Netz jedoch im Laufe seines Trainings und speichert sie in seinen gewichteten Verbindungen.[296]
Dadurch ist die Information in einem Künstlichen Neuronalen Netz dezentral, verteilt auf eine große Zahl von Gewichten gespeichert und dementsprechend schwer zu lokalisieren.[297] Sie ergibt sich aus der Zusammenschau des gesamten Netzes, seiner Struktur aus Neuronen und den zwischen ihnen bestehenden gewichteten Verbindungen, die das Netz durch sein Training erlangt hat:

> „In a neural network of specific architecture, knowledge representation of the surrounding environment is defined by the values taken on by the

[294] *Bermúdez*, Cognitive science: an introduction to the science of the mind, S. 106.
[295] *Zech*, Weizenbaum Series 2020, 1 (28).
[296] *Zech*, Weizenbaum Series 2020, 1 (15).
[297] *Ertel*, Grundkurs Künstliche Intelligenz, S. 308; *Bermúdez*, Cognitive science: an introduction to the science of the mind, S. 141.

> free parameters (i.e., synaptic weights and biases) of the network. The form of this knowledge representation constitutes the very design of the neural network, and therefore holds the key to its performance."[298]

> "What distinguishes the networks are their different patterns of weights. But a pattern of weights is not a rule, or an algorithm of any kind. Rather a particular pattern of weights is what results from the application of one rule (the learning algorithm)."[299]

Dieses "Gewichtsmuster", eingebettet in die Architektur des KNN und seine mathematischen Funktionen,[300] ist mithin die Information des KNN. Sie kann auch als das mathematische Modell der Art und Weise der Informationsverarbeitung im KNN beschrieben werden.[301] Diese Art und Weise der Informationsverarbeitung macht die Bedeutung der Information, ihre semantische Ebene, aus. Denn ein KNN ist ein informationsverarbeitendes System und die Frage, was in diesem System passiert, und auch der wirtschaftliche Wert des Systems liegen in der durch Training ausgebildeten Informationsverarbeitung, also in dessen „Regeln".[302]
Die semantische Information ist jedoch nicht in Form von Zeichen im Netz präsent, sie ist nicht durch Symbole für einen Menschen verständlich codiert.[303] Eine „klare Zuordnung von Zeichen und Bedeutung"[304] fehlt, die es ermöglichen würde, die Informationsverarbeitung im Netz zu analysieren.

---

[298] *Haykin*, Neural Networks: A comprehensive Foundation, S. 47.

[299] *Bermúdez*, Cognitive science: an introduction to the science of the mind, S. 143.

[300] Auch Andrews et. al. zählen die Architektur und die Aktivierungsfunktion zum „Wissen" des Netzes: „[...] within a trained artificial neural network, knowledge acquired during the training phase is encoded as (a) the network architecture itself (e.g. the number of hidden units), (b) an activation function associated with each (hidden and output) unit of the ANN, and (c) a set of (real-valued) numerical parameters (called weights).", *Andrews/Diederich/Tickle*, Knowledge-Based Systems 1995, 373 (375).

[301] Es wird auch von „Entscheidungsstruktur" gesprochen, *Gesellschaft für Informatik*, Technische und rechtliche Betrachtungen algorithmischer Entscheidungsverfahren, S. 54 f.

[302] so wohl auch Zech, nach dem „[d]ie Gesamtheit der Gewichtungen und Schwellenwerte in einem bestimmten Netz [...] implizit die Regeln [repräsentieren], nach denen die Informationsverarbeitung ausgeführt wird", *Zech*, Weizenbaum Series 2020, 1 (43).

[303] vgl. *Zech*, Weizenbaum Series 2020, 1 (33).

[304] *Zech*, Weizenbaum Series 2020, 1 (13).

Wird ein komplexes KNN im Rahmen eines Entscheidungsprozesses eingesetzt, so ist es daher so gut wie unmöglich, den „Lösungsweg" anhand der Gewichte des Netzes nachzuvollziehen.[305] Nicht einmal die Programmierer, die ein hochkomplexes System entworfen haben, können den Grund für die Reaktion des Systems auf eine bestimmte Situation isolieren.[306] Die „symbolische Logik", die klassischen Algorithmen zugrunde liegt, ist nicht mehr gegeben.[307]
Die Information des KNN befindet sich in der beschriebenen „Blackbox" und ist selbst für Experten nicht in vollem Umfang zu erfassen:

> „Selbst wenn wir die zugrunde liegenden mathematischen Prinzipien verstehen, fehlt solchen Modellen eine explizite deklarative Wissensrepräsentation."[308]

Auch durch Größe und Komplexität des Netzes übersteigt die implizit repräsentierte Information die menschliche Auffassungsgabe: Burrell spricht in diesem Zusammenhang von „*opacity as the complexity of scale*".[309] Darin zeigt sich ein Phänomen von Big Data und der modernen Informationsgesellschaft: Information wird nicht mehr nur in Abhängigkeit von einem menschlichen Geist gesehen.[310]
Dass diese Bedeutungsebene für den Menschen nicht verfügbar ist, spricht jedoch nicht gegen einen semantischen Gehalt der Information des KNN. Es handelt sich um potenzielle semantische Information, die erst in dem Moment zu faktischer semantischer Information wird, in dem sie explizit symbolisch repräsentiert wird.

---

[305] *Ertel*, Grundkurs Künstliche Intelligenz, S. 308.
[306] *Knight*, MIT Technology Review 2017, 53 (56).
[307] *Zech*, Weizenbaum Series 2020, 1 (17).
[308] *Holzinger*, Informatik-Spektrum 2018, 138 (138).
[309] *Burrell*, Big Data & Society 2016, 1 (9); siehe zum Verhältnis von Information und Bewusstsein *Zech*, Information als Schutzgegenstand, S. 28 ff.
[310] Siehe zum Phänomen der Abstraktion von Information *Zech*, Information als Schutzgegenstand, S. 167 ff.

B. Schutzmöglichkeiten

Die implizit repräsentierte Information eines KNN, sein „Wissen", macht das Know-how des Netzes aus.[311] Daher ist ihre Schutzfähigkeit von besonderem Interesse und soll im Folgenden untersucht werden.

I. Geschäftsgeheimnisgesetz[312]

Sofern Erlangung und Nutzung eines Geschäftsgeheimnisses nicht den Tatbestand eines der Handlungsverbote des § 4 GeschGehG erfüllen, oder aufgrund anderer Vorschriften verboten sind – etwa durch das Urheber- oder Lauterkeitsrecht – stehen dem Geheimnisinhaber gegen eine Reproduktion seiner geheimen Information keine Ansprüche zu.
Der Geheimnisinhaber hat daher mehrere Interessen: die Offenbarung des Geschäftsgeheimnisses soll verhindert werden, da mit ihr das Geschäftsgeheimnis tatbestandlich verloren geht und die sich aus den §§ 6 ff. GeschGehG ergebenden Ansprüche bei Rechtsverletzungen entfallen. Gleichzeitig soll auch die Erlangung des Geschäftsgeheimnisses durch Dritte vermieden werden, um eine Reproduktion der geheimen Information und mithin eine Schwächung des Wettbewerbsvorteils zu verhindern.
Es ist eine Besonderheit des Geschäftsgeheimnisrechts gegenüber den Immaterialgüterrechten, dass der Schutz als Geheimnis von faktischen Elementen abhängt, weshalb das Geschäftsgeheimnis auch als „unvollkommenes Immaterialgüterrecht" bezeichnet wird.[313]
Ob die Voraussetzungen des § 2 Nr. 1 GeschGehG vorliegen, ist daher in weiten Teilen eine Frage des Einzelfalls, etwa was allgemeine Bekanntheit (§ 2 Nr. 1 lit. a GeschGehG) und Geheimhaltungsmaßnahmen (§ 2 Nr. 1 lit. b GeschGehG) betrifft.

---

[311] nach Hartmann/Prinz ist das „know-how im [trainierten] KNN implizit gespeichert", *Hartmann/Prinz*, in: Taeger, Rechtsfragen digitaler Transformationen - Gestaltung digitaler Veränderungsprozesse durch Recht, S. 783.

[312] Ein Auszug aus dem Geschäftsgeheimnisgesetz mit den hier analysierten Normen findet sich im Anhang.

[313] *Ohly*, in: Harte-Bavendamm/Ohly/Kalbfus, GeschGehG, Einl. Rn. 30 die dogmatische Einordnung des Geschäftsgeheimnisschutzes bleibt auch im neuen Recht umstritten, ist jedoch für die vorliegende Arbeit nicht von Bedeutung. Für einen Überblick siehe ebd., Rn. 21 ff.

Hier soll zum Zwecke der Untersuchung davon ausgegangen werden, dass das KNN nicht als Software auf den Markt gebracht, sondern entweder nur durch den Entwickler bzw. Verwender selbst im Zuge seiner unternehmerischen Tätigkeit genutzt, oder cloudbasiert oder in eine Hardware integriert (*Internet of Things*, IoT) zur Verfügung gestellt wird. Denn würde das KNN als Bestandteil eines Computerprogramms als Software in Verkehr gebracht, so würde, wie bei Verträgen über Software üblich, auch der Maschinencode übergeben.[314] Das könnte dann bereits dazu führen, dass das in ihm enthaltene Geschäftsgeheimnis allgemein bekannt oder zumindest ohne weiteres zugänglich wird.[315]
Die Voraussetzungen der Schutzfähigkeit als Geschäftsgeheimnis bedürfen daher einer differenzierteren Prüfung, auch hinsichtlich der einzelnen Bestandteile der semantischen Information eines KNN.

1. Information eines Künstlichen Neuronalen Netzes als Geschäftsgeheimnis

a) Information

Schutzgegenstand des Geschäftsgeheimnisgesetzes ist semantische Information, unabhängig von ihrer konkreten Verkörperung.[316] Das zeigt sich etwa an den in § 4 Abs. 1 Nr. 1 GeschGehG formulierten Handlungsverboten. Sie knüpfen zwar an unterschiedliche Gegenstände an („Dokumenten, Gegenständen, Materialien, Stoffen oder elektronischen Dateien"), die jedoch nur Träger der semantischen Information sind („die das Geschäftsgeheimnis enthalten oder aus denen sich das Geschäftsgeheimnis ableiten lässt").
Den Schutz semantischer Information hat der Geheimnisschutz mit dem Patentrecht gemein und beide Schutzrechtsregime unterscheiden sich darin vom Urheberrecht, das grundsätzlich nicht die Idee, sondern lediglich die Form und mithin syntaktische Information schützt.[317]
Der Inhalt der Information eines KNN wurde im ersten Teil dieses Kapitels herausgearbeitet: Es sind die Regeln, nach denen die Verarbeitung einer Eingabe erfolgt. Als konkrete Bestandteile dieser Information können maßgeblich die

---

[314] Siehe dazu Redeker, allerdings bezogen auf den Objektcode, der einen Zwischenschritt zwischen Quellcode und Maschinencode darstellt, *Redeker*, IT-Recht, Rn. 204.

[315] Siehe dazu Kapitel 5.

[316] Noch zum alten Recht *Zech*, Information als Schutzgegenstand, S. 231.

[317] *Zech*, Information als Schutzgegenstand, S. 246; *Schulze*, in: Dreier/Schulze: UrhG, § 2 Rn. 37; *Bullinger*, in: Wandtke/Bullinger, UrhG, § 2 Rn. 33 ff.

Netzarchitektur, die verwendeten mathematischen Funktionen und die Gewichte unterschieden werden.
Auch wenn das Geschäftsgeheimnisgesetz auch Ideen schützt, so kann der Schutz dennoch erst beginnen, sobald die Information auf irgendeine Weise verkörpert ist – „Ideen, die lediglich als Gedanken existieren“[318], können nicht als Geschäftsgeheimnisse geschützt werden. Diese Einschränkung besteht grundsätzlich auch für die semantische, implizit im KNN gespeicherte Information, sein „Wissen“. Anders als ein Algorithmus kann die semantische Information eines KNN jedoch nicht lediglich in den Gedanken eines Programmierers existieren. Sie entsteht ja gerade durch das Training des Netzes und ist mithin von Anfang an verkörpert als elektrisches Signal. Daraus folgt, dass sie auch grundsätzlich Information im Sinne des § 2 Nr. 1 GeschGehG darstellen kann.
Auf diesen Gedanken stützt sich die in den folgenden Kapiteln vorgenommene Analyse der unterschiedlichen Darstellungsformen der Information des KNN. Es werden verschiedene Formen syntaktischer Information daraufhin untersucht, in welchem Maße sie die semantische Information des KNN enthalten und daran anknüpfend, ob das Geschäftsgeheimnis durch ihre Offenlegung erlangt wird.
Zunächst muss jedoch die semantische Information des KNN alle Voraussetzungen des § 2 Nr. 1 GeschGehG erfüllen, um als Geschäftsgeheimnis Schutz zu genießen.
Es wurde bereits darauf hingewiesen, dass es sich bei der geschützten Information zumeist um eine Mehrzahl einzelner Informationen handeln wird, die auch jeweils für sich genommen nicht geheim sein müssen („weder insgesamt noch in der genauen Anordnung und Zusammensetzung ihrer Bestandteile [...] allgemein bekannt oder ohne Weiteres zugänglich [...]“, vgl. § 2 Nr. 1 a GeschGehG). In diesem Fall kann die Nichtoffenkundigkeit auf die Kombination dieser Einzelinformationen zurückzuführen sein.
Auch der BGH spricht in seinem Urteil zur SCHUFA ausdrücklich von „als Geschäftsgeheimnis geschützten Inhalten der Scoreformel“ im Plural:

> „Zu den nach dem gesetzgeberischen Willen als Geschäftsgeheimnis geschützten Inhalten der Scoreformel zählen damit die im ersten Schritt in die Scoreformel eingeflossenen allgemeinen Rechengrößen, wie etwa die herangezogenen statistischen Werte, die Gewichtung einzelner Berech-

[318] *Alexander*, in: Köhler/Bornkamm/Feddersen, GeschGehG, § 2 Rn. 25.

> nungselemente bei der Ermittlung des Wahrscheinlichkeitswerts und die Bildung etwaiger Vergleichsgruppen als Grundlage der Scorekarten."[319]

Auch die semantische Information eines KNN setzt sich aus verschiedenen Einzelinformationen zusammen. Die wesentlichen Informationen sind die Netzarchitektur, die Gewichte und Schwellenwerte, sowie die Funktionen (Propagierungs- und Aktivierungsfunktion).
Da die im weiteren Verlauf dieser Arbeit vorgestellten Darstellungsstufen eines KNN diese Einzelinformationen in unterschiedlichem Maße enthalten, soll ihre Schutzfähigkeit als Geschäftsgeheimnis jeweils separat überprüft werden. Die Frage nach der Schutzfähigkeit ihrer Kombination kann darauf aufbauend beantwortet werden. Denn der Geschäftsgeheimnisschutz steht in starker Abhängigkeit zu tatsächlichen Faktoren, die je nach Einzelfall sehr unterschiedlich ausfallen können. Erst eine differenzierte Prüfung sowohl der Einzelinformationen als auch ihrer Kombination ermöglicht dann eine fundierte Bewertung des Geheimnischarakters im Einzelfall.
Die einzelnen Bestandteile des KNN (Netzarchitektur, Gewichte und Schwellenwerte sowie Funktionen) fallen sowohl für sich genommen als auch in Kombination unproblematisch unter den weiten Informationsbegriff des § 2 Nr. 1 GeschGehG.[320] Denn eine Besonderheit des Geschäftsgeheimnisschutzes gegenüber dem Urheber- und Patentrecht ist es gerade, dass auch Ideen, mathematische Modelle, Algorithmen und KI-Modelle als Information im Sinne des § 2 Nr. 1 GeschGehG schutzfähig sind.[321] So fallen auch mathematische Formeln, denen der Schutz durch das Patentrecht versagt bleibt, unter den Informationsbegriff des GeschGehG.[322] Auch das Trainingsergebnis, also die als Datei gespeicherten Gewichte, können grundsätzlich als Geschäftsgeheimnis geschützt sein.[323] Sie

---

[319] BGH, Urteil v. 28.1.2014, VI ZR 156/13, NJW 2014, 1235, Rn. 27.

[320] Siehe zur Information i. S. d. GeschGehG Kapitel 2 A; so explizit auch *Alexander*, in: Köhler/Bornkamm/Feddersen, GeschGehG, § 2 Rn. 26.

[321] Siehe zum Schutz von Algorithmen als Geschäftsgeheimnis BGH, Urteil v. 28.1.2014, VI ZR 156/13, NJW 2014, 1235 (1237); OLG München, MMR 2020, 404 (406); *Scheja*, CR 2018, 485; zum Schutz von KI-Systemen als Geschäftsgeheimnis siehe *Hauck/Cevc*, ZGE 2019, 135 (163 ff.).

[322] Vgl. § 1 Abs. 3 Nr. 1 PatG; siehe zur mathematischen Formel als Geschäftsgeheimnis nach altem Recht LG Stuttgart, NJW 1991, 441 – *Geldspielautomat*.

[323] So auch *Surblyté*, WuW 2017, 120 (125); *Sagstetter*, in: Maute/Mackenrodt, Recht als Infrastruktur für Innovation, S. 294 f.

können unter den weiten Informationsbegriff des § 2 Nr. 1 GeschGehG gefasst werden, der auch computergenerierte Daten einschließt.[324]

b) Allgemeine Bekanntheit

Allgemeine Bekanntheit im Sinne des § 2 Nr. 1a GeschGehG kann im Falle technischer Information nur angenommen werden, wenn sie zum „gängigen Kenntnis- und Wissensstand“[325] des durchschnittlichen Angehörigen des jeweiligen Fachkreises gehört. Das Merkmal ist mithin weiter als im Patentrecht, wo auch schwer zugängliche Information zum Stand der Technik gehören und daher neuheitsschädlich sein kann.[326]
Bei einigen Bestandteilen des Netzes dürfte es sich für sich genommen um gängiges Wissen handeln, das dem durchschnittlichen Programmierer von KNN allgemein bekannt sind. Dabei kommt es jedoch stark auf den Einzelfall an.
So wird die Architektur eines einfachen vorwärtsgerichteten Netzes typischerweise zum gängigen Wissensstand in entsprechenden Fachkreisen gehören. Ein komplizierter aufgebautes, mit unterschiedlichen Rückkoppelungen und Schleifen versehen Netz könnte hingegen schon nicht mehr als „allgemein bekannt“ eingeordnet werden. Denn auch wenn einige Aufbaueigenschaften durch das zu lösende Problem mehr oder weniger vorgegeben sind (Zahl der Eingabe- und Ausgabeneuronen), so erfordert insbesondere die Ermittlung der angemessenen Zahl von verborgenen Schichten und deren Neuronen einen nicht unerheblichen experimentellen Aufwand.[327]
Die Funktionen, die innerhalb der einzelnen Neuronen zum Tragen kommen, insbesondere Aktivierungs- und Propagationsfunktion, dürften hingegen in der Regel allgemein bekannt im beschriebenen Sinne sein. Es gibt wie gezeigt eine begrenzte Anzahl an Funktionen, deren Einsatz in diesem Bereich sinnvoll ist, und ihre Verwendung gehört zum gängigen Wissen eines mit KNN befassten Programmierers.
Anders zu beurteilen ist dieses Merkmal hinsichtlich der Gewichte und Schwellenwerte. Diese sind eine Folge vielzähliger Einzelentscheidungen beim Entwurf des Netzes und Produkt des anschließenden Trainings mit einem spezifischen

[324] Siehe dazu Kapitel 2 A.

[325] *Alexander*, in: Köhler/Bornkamm/Feddersen, GeschGehG, § 2 Rn. 35.

[326] Siehe zum Verhältnis von Nichtoffenkundigkeit im Sinne des GeschGehG und Neuheit im Sinne des PatG *Kalbfus*, in: Harte-Bavendamm/Ohly/Kalbfus, GeschGehG, Einl. Rn. 178.

[327] Siehe dazu oben, Kapitel 1 E.

Datensatz und nach individuellen Trainingsmodalitäten. Sie sind mithin grundsätzlich einzigartig.[328] Daher können sie als in der Regel nicht allgemein bekannt beurteilt werden.
Gleiches muss dann auch für die ein spezielles KNN charakterisierende Kombination aus Netzarchitektur, verwendeten Formeln und Gewichten gelten: sie wird typischerweise den Personen in den Kreisen, die üblicherweise mit dieser Art von Informationen umgehen, nicht allgemein bekannt sein. So wäre zwar ein einfaches vorwärtsgerichtetes Netz mit einer verborgenen Schicht und einer gängigen Aktivierungsfunktion wohl in Fachkreisen allgemein bekannt. Die konkrete Kombination aus Architektur, Funktionen, Gewichten und Schwellenwerten jedoch bietet unzählige Möglichkeiten und ist mithin nicht allgemein bekannt. Denn bereits der Wortlaut des § 2 Nr. 1a GeschGehG („weder insgesamt noch in der genauen Anordnung und Zusammensetzung ihrer Bestandteile") macht deutlich, dass die Geheimhaltung der spezifischen Kombination von Information ausreicht, selbst wenn die einzelnen Bestandteile für sich genommen bekannt sind.[329]

c) Zugänglichkeit ohne weiteres

Sofern die einzelnen Bestandteile der Information des KNN nach den dargestellten Maßstäben nicht generell oder im Einzelfall allgemein bekannt sind, schließt sich die Frage an, ob sie auch nicht ohne weiteres zugänglich sind gemäß § 2 Nr. 1a GeschGehG.
Nicht ohne weiteres zugänglich in diesem Sinne ist die Information, wenn sich die betreffenden Fachkreise nicht ohne große Schwierigkeiten Kenntnis über sie verschaffen können.[330] Von einer leichten Zugänglichkeit wäre auszugehen, wenn eine Verkörperung des Netzes etwa in einem öffentlichen Register einsehbar wäre, oder wenn sie in Verkehr gebracht würde. Die bloße Möglichkeit, sich die Hyperparameter und Gewichte des Netzes durch Techniken des Reverse Engineering zu verschaffen, etwa bei cloudbasierten Anwendungen, reicht hin-

[328] Zum geheimen Charakter neuronaler Netze und von Gewichtsmatrizen siehe auch *Sagstetter*, in: Maute/Mackenrodt, Recht als Infrastruktur für Innovation, S. 294 ff.

[329] Sog. „Mosaiktheorie", *Kalbfus*, WRP 2013, 584 (585); *Ohly*, GRUR 2019, 441 (443); *Harte-Bavendamm*, in: Harte-Bavendamm/Ohly/Kalbfus, GeschGehG, § 2 Rn. 22; *Alexander*, in: Köhler/Bornkamm/Feddersen, GeschGehG, § 2 Rn. 32.

[330] Siehe zu diesem Merkmal eingehend *Alexander*, in: Köhler/Bornkamm/Feddersen, GeschGehG, § 2 Rn. 36.

gegen nicht aus.[331] Dies ergibt sich bereits im Umkehrschluss aus § 3 Abs. 1 Nr. 2 GeschGehG, der die Erlangung des Geschäftsgeheimnisses durch Reverse Engineering erlaubt. Diese Erlaubnis wäre nicht notwendig, wenn die entsprechende Information bereits „ohne weiteres zugänglich" und mithin schon tatbestandlich kein Geschäftsgeheimnis wäre.[332] Auch wenn Daten (etwa die Gewichte) auf fremden Servern oder auf einer Cloud gespeichert werden, ist nicht per se von einer leichten Zugänglichkeit auszugehen.[333] Denn die Betreiber entsprechender Dienste fallen nicht unter den durch § 2 Nr. 1a GeschGehG adressierten Personenkreis.[334] Sie gehen nicht „üblicherweise mit dieser Art von Informationen" um, sondern ihr Geschäftsmodell liegt ganz allgemein im Umgang mit Daten.

Die Frage nach der Zugänglichkeit ohne weiteres lässt sich folglich nur nach rein faktischen Gegebenheiten beantworten. Dementsprechend muss für die zweite Alternative der Geheimhaltung des § 2 Nr. 1a GeschGehG nicht nach den verschiedenen Bestandteilen der Information des KNN differenziert werden. Wird die Funktionalität des Netzes über eine Cloud oder im Rahmen eines IoT-Geräts zur Verfügung gestellt, so sind weder die Einzelinformationen (sofern sie nicht ohnehin allgemein bekannt sind) noch ihre Kombination ohne weiteres zugänglich und mithin geheim im Sinne der Norm. Ebenso wie der Maschinencode ist der Quellcode nicht ohne weiteres zugänglich, solange er nicht mit dem KNN in den Verkehr gebracht wird. Anders als der Maschinencode wird der Quellcode jedoch auch beim Erwerb herkömmlicher, proprietärer Software nicht herausgegeben, da mit ihm bereits umfangreich Piraterie betrieben werden kann.[335]

d) Zwischenfazit: Geheime Information eines Künstlichen Neuronalen Netzes

Sofern insoweit keine Zugänglichkeit ohne Weiteres gegeben ist, ergibt sich hinsichtlich der geheimen semantischen Information eines KNN folgendes Bild: die Gewichte und Schwellenwerte werden typischerweise geheim sein im Sinne des § 2 Nr. 1a GeschGehG, während die verwendeten Funktionen für sich ge-

[331] Siehe zum Reverse Engineering von KNN unten, Kapitel 9.

[332] Siehe zum Ganzen *Alexander*, in: Köhler/Bornkamm/Feddersen, GeschGehG, § 2 Rn. 37 f.

[333] So auch *Sagstetter*, in: Maute/Mackenrodt, Recht als Infrastruktur für Innovation, S. 295.

[334] So auch *Sagstetter*, in: Maute/Mackenrodt, Recht als Infrastruktur für Innovation, S. 295, Fn. 46.

[335] *Söbbing*, CR 2020, 223 (226).

nommen nicht geheim sind. Die Beurteilung der Netzarchitektur erfordert eine Einzelfallbetrachtung, da hier die Komplexität des Aufbaus für die allgemeine Bekanntheit ausschlaggebend ist. Die Kombination dieser Einzelelemente und mithin die semantische Information des KNN in ihrer Gesamtheit jedoch wird typischerweise geheim sein im Sinne des GeschGehG.

e) Wirtschaftlicher Wert

Aus der Geheimhaltung der Information muss sich ein wirtschaftlicher Wert ergeben. Nach Erwägungsgrund 14 der Geschäftsgeheimnis-RL ist das beispielsweise der Fall, wenn die „Offenlegung [des Geheimnisses] die Interessen der Person, die rechtmäßig die Kontrolle über sie ausübt, aller Voraussicht nach dadurch schädigt, dass das wissenschaftliche oder technische Potenzial, die geschäftlichen oder finanziellen Interessen, die strategische Position oder die Wettbewerbsfähigkeit dieser Person untergraben werden."
In die Entwicklung eines KNN fließen erhebliche zeitliche und finanzielle Ressourcen.[336] Dies gilt nicht nur für die Gewichte und Schwellenwerte als Trainingsergebnis, sondern bereits für die Entwicklung einer an die zu lösende Aufgabe angepassten Architektur. Die Möglichkeit, das für die eigenen unternehmerischen Zwecke entwickelte KNN mit seiner spezifischen Funktionalität exklusiv nutzen und anbieten zu können, hat für den Geheimnisträger einen großen wirtschaftlichen Wert und dürfte für seine Wettbewerbsfähigkeit häufig entscheidend sein. Nicht selten würde der Verlust des Geheimnisses das Geschäftsmodell eines Unternehmens in Gefahr bringen.[337] Aus der Geheimhaltung der gesamten semantischen Information des KNN ergibt sich mithin in jedem Fall ein wirtschaftlicher Wert.
Dies könnte allenfalls vor dem Hintergrund in Frage gestellt werden, dass die reine Information über Netzaufbau, Gewichte und Funktionen allein einem etwaigen Konkurrenten nicht notwendigerweise in der Entwicklung eines eigenen Produkts weiterhilft. Häufig dürfte es noch darauf ankommen, ob zusätzlich die Information über das Einsatzgebiet des KNN bekannt ist. Dies wird allerdings in den hier im Fokus stehenden Konstellationen, in denen Unterneh-

[336] *Hartmann/Prinz*, in: Taeger, Rechtsfragen digitaler Transformationen - Gestaltung digitaler Veränderungsprozesse durch Recht, S. 782; *Ehinger/Stiemerling*, CR 2018, 761 (764).
[337] *Apel/Kaulartz*, RDi 2020, 24 (24).

men zur Offenbarung von Informationen verpflichtet werden, regelmäßig der Fall sein.

Auch die Gewichte und Schwellenwerte allein haben aufgrund ihrer Geheimhaltung einen wirtschaftlichen Wert. Denn aus einer Gewichtsmatrix lässt sich die Netzarchitektur ableiten, wodurch die relevante Information des Netzes vorhanden und die Reproduktion eines funktionsgleichen Netzes möglich ist. Ist das Einsatzgebiet bekannt, so kann eine Offenlegung der Gewichte einen bestehenden Wettbewerbsvorteil daher bereits erheblich beeinträchtigen, wenn nicht sogar existenzbedrohend sein.

Ob sich aus der geheimen Netzarchitektur allein ein wirtschaftlicher Wert ergibt, ist hingegen weniger eindeutig zu beantworten. Dies könnte allenfalls vor dem Hintergrund zu bejahen sein, dass die Entwicklung der auf das zu lösende Problem optimal angepassten Architektur bereits einen hohen experimentellen Aufwand erzeugt. Allerdings lässt sich daraus nicht ohne Weiters schließen, dass die Offenlegung der Architektur (als Gegenteil ihrer Geheimhaltung, auf die sich der wirtschaftliche Wert bezieht) auch bereits einen Wettbewerbsnachteil oder eine Schädigung des wissenschaftlichen Potenzials mit sich bringt. Denn der entwicklerische Aufwand muss nicht proportional sein zu dem Nutzen, den ein Konkurrent aus dem Produkt ziehen könnte. Vielmehr muss davon ausgegangen werden, dass aus der Geheimhaltung der, wenn auch komplexen, Netzarchitektur allein nicht grundsätzlich ein wirtschaftlicher Wert folgt. Es wird im Einzelfall eher darauf ankommen, ob sie besondere Merkmale aufweist, die einen Wettbewerbsvorteil darstellen. Bedeutung wird darüber hinaus der Frage zukommen, ob der Einsatzbereich des Netzes bekannt ist oder sich durch den Wettbewerber ableiten oder zumindest eingrenzen lässt.[338] Dann könnte nämlich die Architektur allein bereits das Training mit eigenen Daten ermöglichen und so eine abgekürzte Entwicklung eines entsprechenden KNN ermöglichen. Auch vor diesem Hintergrund könnte ein wirtschaftlicher Wert aufgrund von Geheimhaltung zu bejahen sein.

Der wirtschaftliche Wert wird mithin für die Kombination der geheimen Einzelinformationen des KNN regelmäßig zu bejahen sein, ebenso für die Gewichtungen und Schwellenwerte. Hinsichtlich der Netzarchitektur ist wiederum eine differenziertere Betrachtung im Einzelfall erforderlich. Zumindest für

[338] Bei einer Ausgabeschicht mit 26 Neuronen könnte etwa auf den Zweck der Buchstabenerkennung geschlossen werden.

komplexere Netze, deren Einsatzbereich bekannt ist, ist regelmäßig von einem wirtschaftlichen Wert auszugehen sein.

f) Angemessene Geheimhaltungsmaßnahmen

Die semantische Information des Netzes muss darüber hinaus nach § 2 Nr. 1b GeschGehG Gegenstand von den Umständen nach angemessenen Geheimhaltungsmaßnahmen sein.
Welche Maßnahmen angemessen sind, ergibt sich weder unmittelbar aus dem GeschGehG noch aus der GeschGeh-RL.[339] Es kann jedoch angenommen werden, dass Schutzmaßnahmen organisatorischer, personeller, technischer und juristischer Art sein sollten.[340] Dabei können auch branchenspezifische Besonderheiten eine Rolle spielen: sind bestimmte Unternehmen häufiges Ziel von Angriffen auf ihre Unternehmensinhalte, so muss das bei der Beurteilung der Angemessenheit von Schutzmaßnahmen berücksichtigt werden.[341] Allgemein gilt, dass das Schutzsystem mit Sorgfalt eingerichtet, seine Einhaltung überwacht und die Aktualität der Maßnahmen regelmäßig überprüft werden müssen.[342] An die Angemessenheit der Maßnahmen sind allerdings keine zu hohen Anforderungen zu stellen, keineswegs sind optimale Schutzmaßnahmen zu fordern.[343] Das erforderliche Maß der Vorkehrungen hängt vielmehr vom Wert der Geheimhaltung ab.[344] Das zeigt auch die Begründung des Gesetzgebers, wonach bei

---

[339] Siehe für eine eingehende Analyse des Merkmals *Maaßen*, GRUR 2019, 352.

[340] Harte-Bavendamm spricht von einem „Vorkehrungsmix", *Harte-Bavendamm*, in: Harte-Bavendamm/Ohly/Kalbfus, GeschGehG, § 2 Rn. 55; die Begründung zum Entwurf des GeschGehG nennt physische und vertragliche Vorkehrungen, BT-Drs. 19/4724, S. 24; siehe auch die ausführliche Maßnahmenumschreibung bei *Huber*, in: Ann/Loschelder/Grosch, Praxishandbuch Know-how-Schutz, S. 595 ff.

[341] Siehe dazu *Hauck/Cevc*, ZGE 2019, 135 (164 ff.).

[342] *Harte-Bavendamm*, in: Harte-Bavendamm/Ohly/Kalbfus, GeschGehG, § 2 Nr. 42.

[343] *Kalbfus*, GRUR-Prax 2017, 391 (392); *Wiese*, Die EU-Richtlinie über den Schutz vertraulichen Know-hows und vertraulicher Geschäftsinformationen, S. 51; *Maaßen*, GRUR 2019, 352 (353).

[344] *Harte-Bavendamm*, in: Harte-Bavendamm/Ohly/Kalbfus, GeschGehG, § 2 Rn. 59; Zur Bewertung der Angemessenheit der Geheimhaltungsmaßnahmen speziell im Bereich von Industrie 4.0 wird vorgeschlagen, den Maßstab des § 8a BSIG heranzuziehen, *Müllmann*, WRP 2018, 1177 (1182). Vor dem Hintergrund, dass an die Angemessenheit der Geheimhaltungsmaßnahmen keine zu hohen Anforderungen zu stellen sind, erscheint eine Orientierung an den für Betreiber Kritischer Infrastrukturen geltenden Maßstäben jedoch zu hoch gegriffen. Denn

der Bewertung der Angemessenheit insbesondere folgende Faktoren eine Rolle spielen:

> „der Wert des Geschäftsgeheimnisses und dessen Entwicklungskosten, die Natur der Informationen, die Bedeutung für das Unternehmen, die Größe des Unternehmens, die üblichen Geheimhaltungsmaßnahmen in dem Unternehmen, die Art der Kennzeichnung der Informationen und vereinbarte vertragliche Regelungen mit Arbeitnehmern und Geschäftspartnern."[345]

Zur Festlegung der jeweils erforderlichen Schutzmaßstäbe werden daher häufig nach einem Drei-Stufen-Modell verschiedene Arten von Know-how unterschieden: wirkliches „Schlüssel-Know-how", strategisch sehr relevantes Know-how und sonstiges wirtschaftlich relevantes Know-how.[346]
Auf der ersten Stufe anzusiedeln sind Informationen, bei deren Verlust der Inhaber in seiner Existenz gefährdet wäre; der Verlust von Know-how der zweiten Stufe würde sich in erheblichem Maße auf die Wettbewerbsfähigkeit auswirken.[347] Gehen Informationen der dritten Stufe verloren, wäre das kurzfristig ökonomisch nachteilig.[348] Aus alledem kann geschlussfolgert werden: „Die „Kronjuwelen" des Geheimwissens, [...], verlangen und verdienen besonders aufwändige Schutzmaßnahmen." [349]
Ein KNN beziehungsweise seine semantische Information insgesamt kann, je nach Unternehmensschwerpunkt des Geheimnisinhabers, mit Fug und Recht als die „Kronjuwelen" des Geheimwissens und mithin als Schlüssel-Know-how bezeichnet werden. Sein Schutz bedarf daher besonderer Vorkehrungen, wie Verschlüsselung, Passwortschutz, eingeschränkte Zugriffsrechte der Mitarbeiter und Verschwiegenheitsvereinbarungen.[350] Dies gilt in gesteigertem Maße, wenn

---

anders als der Schutz von Geheimnissen gewöhnlicher Unternehmen dient der Schutz der Informationstechnik Kritischer Infrastrukturen nicht nur dem Unternehmen selbst, sondern zuvorderst dem Gemeinwesen durch Schutz vor Versorgungsengpässen und vor Gefährdungen der öffentlichen Sicherheit (vgl. § 2 Abs. 10 Nr. 2 BSIG).

[345] BT-Drs. 19/4724, S. 24 f.

[346] *Kalbfus*, GRUR-Prax 2017, 391 (393); *Maaßen*, GRUR 2019, 352 (356).

[347] *Kalbfus*, GRUR-Prax 2017, 391 (393).

[348] *Maaßen*, GRUR 2019, 352 (356).

[349] *Harte-Bavendamm*, in: Harte-Bavendamm/Ohly/Kalbfus, GeschGehG, § 2 Rn. 59.

[350] Einen Maßnahmenkatalog zum Schutz von Algorithmen als Geschäftsgeheimnis entwirft *Scheja*, CR 2018, 485 (490 ff.); siehe zu weiteren Maßnahmen *Huber*, in: Ann/Loschel-

das KNN einem Dritten zur Nutzung übergeben wird. Denn zur Wahrung des Geheimnisses in Bezug auf Externe darf grundsätzlich nur die unbedingt notwendige Information herausgegeben werden (sog. need-to-know Konzept).[351] Soll daher ausnahmsweise eine Verkörperung des KNN herausgegeben werden, etwa der Maschinencode, kann nur durch wirksame Vertraulichkeitsvereinbarungen die Geheimhaltung gewahrt bleiben, die zudem wohl individualvertraglich festgelegt werden müssen.[352] Wird die Funktionalität des KNN cloudbasiert angeboten, so sind besondere technische Schutzmaßnahmen unabdinglich.[353] In diesem Zusammenhang können *Software-as-a-Service*-Techniken einen höheren Schutz bieten, bei denen das KNN im System des Geheimnisinhabers bleibt.[354] Damit kann auch das Risiko einer Rückwärtsanalyse abgemildert werden, das ebenfalls bei der Beurteilung des angemessenen Schutzes bedacht werden sollte.[355]

Auch die Gewichte für sich genommen müssen durch Geheimhaltungsmaßnahmen geschützt sein, die ihrer Bedeutung als Schlüssel-Know-how gerecht werden. Darunter wird man wohl verstehen können, dass die entsprechende Datei verschlüsselt ist, dass nur ein beschränkter Personenkreis Zugriff auf sie hat und dass dafür Sorge getragen wird, dass Mitarbeiter über ihre Verschwiegenheitsverpflichtungen belehrt und nach Möglichkeit auch bei ihrem Ausscheiden an diese erinnert werden.[356]

---

der/Grosch, Praxishandbuch Know-how-Schutz, S. 595 ff.; *Alexander*, in: Köhler/Bornkamm/Feddersen, GeschGehG, § 2 Rn. 53 ff.; *Harte-Bavendamm*, in: Harte-Bavendamm/Ohly/Kalbfus, GeschGehG, § 2 Rn. 40 ff.

[351] *Harte-Bavendamm*, in: Harte-Bavendamm/Ohly/Kalbfus, GeschGehG, § 2 Rn. 64, m.w.N.

[352] Siehe zur Frage der angemessenen Geheimhaltungsmaßnahmen bei KNN *Kuß/Sassenberg*, in: Sassenberg/Faber, Rechtshandbuch Industrie 4.0 und Internet of Things: Praxisfragen und Perspektiven der digitalen Zukunft, S. 449, die sich jedoch auf „Algorithmus und weights" beziehen, nicht auf den Maschinencode.

[353] Ann bezeichnet die IT-Sicherheit als „lückenhaft", wenn kritische Daten in eine Cloud geladen werden, *Ann*, GRUR 2014, 12 (14).

[354] *Apel/Kaulartz*, RDi 2020, 24 (31).

[355] So auch *Apel/Kaulartz*, RDi 2020, 24 (31).

[356] Sog. Exit-Interviews; siehe dazu und allgemein zu betriebsinterner Absicherung mit Blick auf Mitarbeiter *Harte-Bavendamm*, in: Harte-Bavendamm/Ohly/Kalbfus, GeschGehG, § 2 Rn. 61 ff.

g) Berechtigtes Interesse an der Geheimhaltung

Nach § 2 Nr. 1c GeschGehG muss darüber hinaus ein berechtigtes Interesse an der Geheimhaltung des KNN bestehen. Die Reichweite dieses Merkmals ist, wie schon nach alter Rechtslage, umstritten. Im Kern geht es um die Frage, ob auch Informationen über rechtswidrige Vorgänge dem Schutz als Geschäftsgeheimnis unterliegen können.[357] Für das KNN als Schutzgegenstand des Geheimnisschutzes kann sich diese Frage etwa stellen, wenn bei seinem Training datenschutzrechtliche Vorschriften missachtet wurden. Teilweise wird Informationen, die auf einer datenschutzwidrigen Erhebung oder Verarbeitung von Daten beruhen, ein berechtigtes Interesse an der Geheimhaltung abgesprochen.[358] Andere halten die Einschränkung des § 2 Nr. 1c GeschGehG für unbeachtlich und gehen davon aus, dass gegebenenfalls bestehenden überwiegenden Interessen der Allgemeinheit oder von Einzelpersonen an einer Offenlegung durch § 3 Abs. 2 und § 5 GeschGehG Rechnung getragen werden könne.[359] Ob ein berechtigtes Geheimhaltungsinteresse am KNN besteht, ist mithin eine Frage des Einzelfalls. Sofern es nicht auf einer rechtswidrigen Datenverarbeitung beruht, ist selbst nach einer restriktiven Auslegung ein Geheimhaltungsinteresse jedoch zu bejahen.
Bei der genauen Betrachtung der Informationsbestandteile ergibt sich daher folgendes: Sofern Gewichte und Schwellenwerte nicht das Ergebnis einer rechtswidrigen Datenverarbeitung durch das KNN sind, ist ein Geheimhaltungsinteresse zu bejahen. Beim Entwurf der Netzarchitektur kommt es üblicherweise nicht zu einer Verarbeitung personenbezogener Daten, die, sofern sie rechtswidrig erfolgt, eine mögliche Einschränkung des berechtigten Interesses darstellen könnte. Für die Netzarchitektur eines KNN ist daher grundsätzlich von einem berechtigten Geheimhaltungsinteresse auszugehen.

h) Zusammenfassung

Zusammenfassend lässt sich festhalten, dass die Netzarchitektur sowie Gewichte und Schwellenwerte – in Abhängigkeit von den dargestellten rechtlichen und faktischen Bedingungen – grundsätzlich als Geschäftsgeheimnis geschützt

---

[357] Siehe zur Auslegung des Merkmals, auch nach altem Recht *Alexander*, in: Köhler/Bornkamm/Feddersen, GeschGehG, § 2 Rn. 78 ff.; *Harte-Bavendamm*, in: Harte-Bavendamm/Ohly/Kalbfus, GeschGehG, § 2 Rn. 66 ff.

[358] *Alexander*, in: Köhler/Bornkamm/Feddersen, GeschGehG, § 2 Rn. 80.

[359] *Harte-Bavendamm*, in: Harte-Bavendamm/Ohly/Kalbfus, GeschGehG, § 2 Rn. 69.

sein können. Besonderer Bedeutung kommt dabei der Angemessenheit von Geheimhaltungsmaßnahmen zu. Lediglich die verwendeten mathematischen Funktionen dürften nicht geheim und mithin für sich genommen nicht schutzfähig sein.
Dies verhindert jedoch nicht, dass auch die semantische Information des KNN als Ganzes ein Geschäftsgeheimnis darstellen kann, da die Einzelinformationen für sich genommen nicht schutzfähig sein müssen. Auch das KNN in seiner Gesamtheit fällt daher grundsätzlich unter den Schutz des GeschGehG.

2. Erlangung des Geschäftsgeheimnisses an einem trainierten Künstlichen Neuronalen Netz

Ein Geschäftsgeheimnis kann rein tatsächlich, unabhängig von einer rechtlichen Bewertung als befugt oder unbefugt, auf unterschiedliche Weise erlangt werden. Eine nicht abschließende Aufzählung von Möglichkeiten kann § 3 Abs. 1 und § 4 Abs. 1 GeschGehG entnommen werden. So ist eine Erlangung etwa möglich durch

> „Beobachten, Untersuchen, Rückbauen oder Testen eines Produkts oder Gegenstands" (§ 3 Abs. 1 Nr. 2a)

oder durch

> „Zugang zu, [...] Aneignung oder [...] Kopieren von Dokumenten, Gegenständen, Materialien, Stoffen oder elektronischen Dateien, [...] die das Geschäftsgeheimnis enthalten oder aus denen sich das Geschäftsgeheimnis ableiten lässt" (§ 4 Abs. 1 Nr. 1 GeschGehG).

Diese Beispiele zeigen, dass das Geschäftsgeheimnis in ganz unterschiedlichen Gegenständen und Formen enthalten sein oder aus diesen abgeleitet werden kann. Dementsprechend wichtig für den Geheimnisschutz ist die genaue Analyse, in welchem Maße eine bestimmte Darstellungsform[360] das Geschäftsgeheimnis am KNN enthält beziehungsweise inwieweit es sich aus dieser ableiten lassen kann. Denn die semantische Information des KNN kann in den einzelnen

[360] Hier wird allgemeiner von „Darstellungsform" und nicht von „Gegenstand" gesprochen, da es nicht auf den Gegenstand (im Sinne struktureller Information) ankommt, sondern auf die syntaktische Information. Ob etwa der Quellcode eines KNN als elektronische Datei oder als Ausdruck auf Papier vorliegt, ist aus Sicht des Geheimnisschutzes irrelevant.

Darstellungsstufen auch nur teilweise enthalten sein. So lassen sich etwa aus dem Quellcode des KNN allein die Gewichte und mithin ein bedeutender Teil seiner Information nicht ableiten – die Architektur des Netzes jedoch, die einen anderen wichtigen Baustein der semantischen Information des Netzes bildet, lässt sich aus dem Quellcode sehr wohl ablesen.[361]

Dementsprechend kann die Erlangung auch einer Teilinformation des KNN durch einen Dritten für den Geheimnisinhaber je nach Informationsgehalt vollkommen unbedenklich oder höchst riskant mit Blick auf die Erlangung weiterer Teilelemente und damit der gesamten Information des KNN sein. Der Informationsgehalt jeder Darstellungsform und das daran anknüpfende Risiko für den Geheimnisinhaber müssen dann auch in die Abwägung im Rahmen von Transparenzpflichten einfließen.

An diesem Punkt stellt sich die Frage, ab welchem Informationsgrad die Information des KNN überhaupt erlangt ist.

Die als Geschäftsgeheimnis geschützt semantische Information des KNN besteht in den Regeln der Informationsverarbeitung des Netzes.[362] Dieser Fokus auf die Regeln der Informationsverarbeitung hat eine wichtige Konsequenz im Hinblick auf die Erlangung des Geschäftsgeheimnisses: Das Geschäftsgeheimnis ist nach hiesigem Verständnis erlangt, wenn ein funktionsgleiches Netz nachgebaut werden kann, das heißt ein Netz, das auf eine Eingabe hin exakt die gleiche Ausgabe produziert wie das Original. Dieses Netz kann sich dann zwar möglicherweise in Details der Architektur und Gewichtungen vom Original unterscheiden. Dennoch enthält es dieselbe semantische Information, nämlich die Regeln der Informationsverarbeitung des Originals.

Nach diesem Verständnis überschneiden sich hier die Offenlegung des Geschäftsgeheimnisses mit der Ausführbarkeit im Patentrecht (Art. 83 EPÜ). Diese ist gegeben, wenn

> „der Durchschnittsfachmann auf Grund der in der Anmeldung enthaltenen Informationen in der Lage ist, unter Inanspruchnahme des von ihm zu erwartenden Informations- und Wissensstandes und des allgemeinen Fachwissens und mit Hilfe der vom Anmelder aufgezeigten Ausführungswege die Lehre zum technischen Handeln, die in den Anmeldungsunterlagen beschrieben und beansprucht ist, zuverlässig, wiederholbar

[361] Siehe dazu Kapitel 6.

[362] Siehe hierzu oben unter A.

> und ohne Umwege in die Praxis umzusetzen, ohne dabei einen unzumutbaren Aufwand treiben und eine unangemessene Zahl anfänglicher Fehlschläge (undue experimentation) hinnehmen zu müssen."[363]

Hinsichtlich der Genauigkeit ist von besonderem Interesse, dass die Ausführbarkeit einer Erfindung bereits dann bejaht wird, wenn ein Fachmann die Erfindung „in praktisch ausreichendem Maße"[364] nachbauen kann, ohne dass es auf eine exakte Wiederholbarkeit ankommt.[365] Der Blick auf das Patentrecht kann somit als Orientierungshilfe dienen, anhand welcher Information der Nachbau eines funktionsgleichen Netzes möglich und das Geschäftsgeheimnis mithin erlangt sein kann.[366]

Für die konkretesten Darstellungsstufen der Information eines KNN, die Stufen 1 und 2, sind die Maßstäbe des Patentrechts allerdings noch nicht von Bedeutung. Wer in Besitz des Maschinencodes oder von Quellcode und Gewichten gelangt, der besitzt die gesamte semantischen Information des KNN, kann es reproduzieren und erlangt mithin das Geschäftsgeheimnis. Dabei spielt es keine Rolle, ob es sich um potenzielle, also implizit repräsentierte, oder faktische, also explizit repräsentierte Information handelt. Erlangt eine Person Besitz etwa an der Binärdatei eines KNN, so kann sie das Netz durch einfaches Kopieren der Datei nacharbeiten, ohne dass die Informationsverarbeitung des Netzes jemals explizit repräsentiert gewesen und für den menschlichen Anwender intellektuell verfügbar gewesen wäre.

Interessant wird die Betrachtung der Reproduzierbarkeit bei den abstrakteren Darstellungsstufen, den Stufen 3 und 4. Hier kann unter Zuhilfenahme der patentrechtlichen Maßstäbe Reproduzierbarkeit angenommen werden, wenn ein Fachmann anhand der offengelegten Information mit zumutbarem Aufwand ein funktionsgleiches KNN nachbauen könnte.

---

[363] *Schäfers/Wieser/Kinkeldey*, in: Benkard, EPÜ, Art. 83 Rn. 62.

[364] BGH, GRUR 62, 80 (81) – *Rohrdichtung*.

[365] EPA, Beschl. v. 27.1.1988, T 0281/86 – *Präprothaumatin*.

[366] Dem Blick auf das Patentrecht als Maßstab für Erlangung des Geschäftsgeheimnisses durch Reproduzierbarkeit steht nicht entgegen, dass das Patentrecht die Ausführbarkeit durch einen Fachmann ausreichen lässt. Da hiesigen Erachtens auch potenzielle semantische Information als Geschäftsgeheimnis geschützt sein kann, kann es zur Erlangung nicht auf den Empfängerhorizont ankommen. Jedenfalls einem Fachmann muss ein Nachbau möglich sein - dass ein Laie ein funktionsgleiches Netz nicht nachbauen kann, ist insofern kein Hindernis.

Die Reproduktion eines funktionsgleichen Netzes setzt grundsätzlich die Kenntnis der semantischen Information des Netzes voraus, also der Netzarchitektur, der Gewichte und Schwellenwerte sowie der verwendeten Funktionen. Zum erfolgreichen Einsatz wird außerdem meist auch das Wissen um die Datenaufbereitung notwendig sein. [367]

Dies gilt jedoch nicht ausnahmslos: Mit Techniken des *Reverse Engineering* kann ein – je nach Technik vollständig oder beinahe – funktionsgleiches ML-Modell reproduziert werden, ohne dass dafür die implizit repräsentierte Information und mithin Funktionsweise des kopierten Netzes entschlüsselt werden müsste.[368]

Das Geschäftsgeheimnisgesetz erlaubt in § 3 die Erlangung von Geschäftsgeheimnissen unter bestimmten Umständen. Nach § 3 Abs. 1 Nr. 2 GeschGehG ist daher etwa explizit das soeben beschriebene *Reverse Engineering* erlaubt. Ebenso ist die eigenständige Schöpfung oder Entdeckung (Nr. 1) und die Erlangung durch Ausübung von Rechten der Arbeitnehmer oder der Arbeitnehmervertretung (Nr. 3) gestattet.

Die Vorschrift des § 3 Abs. 2 GeschGehG statuiert eine Selbstverständlichkeit:[369] die Erlangung eines Geschäftsgeheimnisses ist über die Fälle des Absatzes 1 hinaus erlaubt, wenn dies durch Gesetz, aufgrund eines Gesetzes oder durch Rechtsgeschäft gestattet ist. An der Schnittstelle von gesetzlichen Transparenzpflichten und Geheimnisschutz ist die gesetzlich erlaubte Erlangung von Geschäftsgeheimnissen naturgemäß von Interesse. Die Vorschrift bedeutet jedoch nicht, dass gesetzliche Vorschriften, welche die Erlangung von Geschäftsgeheimnissen erlauben, dem GeschGehG grundsätzlich vorgehen. Zumindest sofern diese Vorschriften eine Interessenabwägung vorsehen oder Ermessen einräumen, müssen Geschäftsgeheimnisse weiterhin berücksichtigt werden. Der Ausgleich zwischen dem jeweiligen Informations- oder Auskunftsrecht und dem Geheimnisschutz findet daher nicht über § 3 Abs. 2 GeschGehG statt, sondern über den Rechtstext, der die Transparenzpflicht vorsieht.[370] Bei der Beant-

[367] Die Vorbereitung der Eingabedaten ist nach Lyre oft ausschlaggebend für ein erfolgreiches Netz, *Lyre*, Informationstheorie. Eine philosophisch-naturwissenschaftliche Einführung, S. 154.

[368] Siehe zum Reverse Engineering und zu diesbezüglichen geheimnisschutzrechtlichen Fragen Kapitel 9.

[369] *Ohly*, in: Harte-Bavendamm/Ohly/Kalbfus, GeschGehG, § 3 Rn. 47.

[370] Siehe zum Ganzen *Ohly*, in: Harte-Bavendamm/Ohly/Kalbfus, GeschGehG, § 3 Rn. 47 ff.

wortung der bereits beschriebenen Frage, wie ein Ausgleich von Geheimnisschutz und Transparenz im Rahmen von DSGVO und KI-VO-E hergestellt werden soll, vermag § 3 Abs. 2 GeschGehG daher nicht weiterzuhelfen.
Das Erlangen des Geschäftsgeheimnisses alleine führt jedoch noch nicht zu dessen Verlust. Denn das Geschäftsgeheimnis besteht so lange, wie die Voraussetzungen nach § 2 Nr. 1 GeschGehG vorliegen, also insbesondere mangelnde allgemeine Bekanntheit oder Zugänglichkeit und angemessene Schutzmaßnahmen.[371] Erlangt etwa ein Konkurrent das Geschäftsgeheimnis an einem KNN und behält er die Information für sich, so bleibt der Geheimnischarakter gewahrt.[372] Allerdings wächst mit dem Erlangen durch Dritte die Gefahr des Bekanntwerdens oder Zugänglichwerdens und damit des Verlusts des Geschäftsgeheimnisses. Auch dieses Risiko wird in der Betrachtung der verschiedenen Darstellungsformen berücksichtigt werden.
In den nachfolgenden Kapiteln wird daher jeweils dargestellt, in welchem Maße eine Darstellungsform des KNN dessen gesamte semantische Information enthält. Dies bemisst sich nach der Reproduzierbarkeit des Netzes anhand der Darstellungsform. Dann wird geprüft, ob die Information explizit wird, ob sich die Darstellungsform also zur Erfüllung von Transparenzpflichten gegenüber Laien eignet.
Diese differenzierte Prüfung ermöglicht die Beantwortung der Frage, die Offenlegung welcher Information über das KNN für den Geheimnisinhaber riskant und welcher eher unbedenklich ist. Dieses Ergebnis kann dann die Abwägung zwischen Transparenzpflichten und Geheimnisschutz erheblich erleichtern.
Nicht geprüft wird im Folgenden, inwiefern die verschiedenen Darstellungsformen selbst als Geschäftsgeheimnis geschützt sind. Sie werden vielmehr nur als Träger der semantischen Information des KNN betrachtet. Mögliche zusätzliche Information, die in ihnen enthalten sein und ebenfalls ein Geschäftsgeheimnis darstellen könnte, wird nicht untersucht.

---

[371] *Alexander*, in: Köhler/Bornkamm/Feddersen, GeschGehG, GeschGehG, § 1 Rn. 17.

[372] *Hoppe*, in: Hoppe/Oldekop, Geschäftsgeheimnisse. Schutz von Know-how und Geschäftsinformationen. Praktikerhandbuch mit Mustern, S. 104.

3. Handlungsverbote und Ansprüche bei Rechtsverletzung

Um das Risiko einer Offenlegung des Geschäftsgeheimnisses am KNN abschätzen zu können, müssen auch die durch das GeschGehG eröffneten Handlungsoptionen des Geheimnisinhabers in den Blick genommen werden.
Verbotene Handlungen in Bezug auf Geschäftsgeheimnisse sind in § 4 GeschGehG abschließend normiert.
Die unbefugte Erlangung gemäß § 4 Abs. 1 GeschGehG ist für die vorliegend betrachteten Fallkonstellation nicht von Relevanz, da das Geschäftsgeheimnis, das zur Erfüllung einer Transparenzverpflichtung übergeben wird, gerade nicht unbefugt (Nr. 1) oder treuwidrig (Nr. 2) erlangt wird.
Von besonderer Bedeutung sind dagegen die Tatbestände der verbotenen Nutzung und Offenlegung nach § 4 Abs. 2 Nr. 2 und 3 GeschGehG. Danach darf ein Geschäftsgeheimnis nicht nutzen oder offenlegen, wer gegen eine Verpflichtung zur Beschränkung der Nutzung des Geschäftsgeheimnisses verstößt (Nr. 2) oder wer gegen eine Verpflichtung verstößt, das Geschäftsgeheimnis nicht offenzulegen (Nr. 3). Dies betrifft mithin die Fälle, in denen zur Erfüllung einer Transparenzverpflichtung eine Offenlegung gegenüber Privaten erfolgt und der Inhaber des Geschäftsgeheimnisses dieses durch eine Vertraulichkeitsvereinbarung abzusichern sucht. Legt der Empfänger der geheimen Information diese in der Folge offen oder nutzt er sie, so begeht er damit eine Rechtsverletzung, vorbehaltlich des Vorliegens eines Ausnahmetatbestandes gemäß § 5 GeschGehG.
Nutzen im Sinne der Norm meint jede wirtschaftlich relevante Verwertung des Geheimnisses,[373] während Offenlegen die „Eröffnung des Geschäftsgeheimnisses gegenüber Dritten, nicht notwendigerweise der Öffentlichkeit“[374], meint.
Bezogen auf das Modell eines trainierten Künstlichen Neuronalen Netzes könnte die wirtschaftliche Verwertung etwa im Anbieten der Funktionalität des

[373] So die einstimmige Definition in der Literatur: *Hauck/Kamlah*, in: MüKo Lauterkeitsrecht, Band 2, § 4 GeschGehG Rn. 22; *Ohly*, in: Harte-Bavendamm/Ohly/Kalbfus, GeschGehG, § 4 Rn. 27; auch wenn der Gesetzentwurf weiter von „jede[r] Verwendung des Geschäftsgeheimnisses“ spricht: Entwurf eines Gesetzes zur Umsetzung der Richtlinie (EU) 2016/943 zum Schutz von Geschäftsgeheimnissen vor rechtswidrigem Erwerb sowie rechtswidriger Nutzung und Offenlegung, BT-Drs. 19/4724, S. 27.

[374] Entwurf eines Gesetzes zur Umsetzung der Richtlinie (EU) 2016/943 zum Schutz von Geschäftsgeheimnissen vor rechtswidrigem Erwerb sowie rechtswidriger Nutzung und Offenlegung, BT-Drs. 19/4724, S. 27.

Netzes über eine eigene Anwendung bestehen, die Offenlegung beispielsweise im Hochladen des Maschinencodes oder des Quellcodes in ein Internetforum. Gegen derartige Handlungen könnte sich der Inhaber des Geschäftsgeheimnisses mit den in den §§ 6 ff. GeschGehG normierten Ansprüchen zur Wehr setzen. Er könnte Beseitigung oder Unterlassung der Rechtsverletzung verlangen (§ 6 GeschGehG), die Vernichtung oder Herausgabe von Verkörperungen des Geschäftsgeheimnisses sowie Rückruf, Entfernung, Vernichtung oder Rücknahme vom Markt eines rechtsverletzenden Produkts verlangen (§ 7 GeschGehG), wobei sämtliche Ansprüche unter dem Vorbehalt der Unverhältnismäßigkeit (§ 9 GeschGehG) und des Missbrauchs (§ 14 GeschGehG) stehen. Flankiert werden die Ansprüche des Geheimnisinhabers von schadenersatzbewährten Auskunftsansprüchen (§ 8 GeschGehG) sowie von Ansprüchen auf Schadenersatz und Abfindung (§§ 10, 11 GeschGehG). Unter bestimmten Voraussetzungen kann sogar eine Strafbarkeit des Rechtsverletzers gegeben sein gemäß § 23 GeschGehG.
All diese Handlungsmöglichkeiten bei Rechtsverletzung ändern nichts an der Tatsache, dass der Geheimnisschutz mit der faktischen Geheimhaltung der Information steht und fällt. Auch wenn dem Inhaber des Geschäftsgeheimnisses mithin die genannten Ansprüche (auch auf Schadenersatz) gegen einen Rechtsverletzer zustehen, so sind sie mit den üblichen Schwierigkeiten der Beweisführung und Rechtsdurchsetzung verbunden. Stellt das betroffene Geschäftsgeheimnis den Kern des Geschäftsmodells des Geheimnisinhabers dar, so werden die gesetzlich normierten Ansprüche, sofern sie sich denn durchsetzen lassen, schwerlich die existenzbedrohende Wirkung eines Geheimnisverlusts abfangen können.

II. Urheberrecht

Anders als im Geheimnisschutzrecht genießen im Urheberrecht Ideen und Grundsätze, die einem Computerprogramm zugrunde liegen, keinen Schutz (§ 69a Abs. 2 S. 2 UrhG). Auch wenn es sich bei der Information des KNN strenggenommen nicht um eine Idee handelt, da sie gerade nicht nur im menschlichen Geist existieren kann, so gilt dieser Grundsatz doch auch für die semantische Information des KNN. Sie ist als Werkinhalt nicht schutzfähig, lediglich ihre verschiedenen Ausdrucksformen können Schutz genießen (§ 69a Abs. 2 S. 1 UrhG).

Der Schutz als Computerprogramm würde darüber hinaus voraussetzen, dass ein individuelles Werk das Ergebnis einer eigenen geistigen Schöpfung ist. Die Schutzvoraussetzungen sind mithin, wie bei anderen Werken auch, „persönliche Schöpfung, geistiger Gehalt, Formgestaltung und Individualität."[375] Das Prinzip der Formgestaltung besagt, dass das Werk eine für menschliche Sinne wahrnehmbare Ausgestaltung angenommen haben muss, wobei es genügt, dass das Netz digital auf einem Datenträger gespeichert ist.[376] Daraus folgt jedoch auch, dass die urheberrechtliche Schutzfähigkeit des KNN nur anhand seiner Ausdrucksformen im Sinne des § 69a Abs. 2 S. 1 UrhG beurteilt werden kann, also nur in Bezug auf seine syntaktische Information. Die urheberrechtliche Schutzfähigkeit dieser Information eines KNN wird daher erst in den folgenden Kapiteln geprüft.

III. Patentrecht

Beim Inhalt der Information eines KNN handelt es sich um die Regeln der Informationsverarbeitung im Netz und damit um Berechnungsvorschriften. Sie sind daher schon mangels Technizität keine Erfindungen auf einem Gebiet der Technik gemäß § 1 Abs. 1 PatG und für sich genommen nicht patentierbar. Der Ausschlusstatbestand des § 1 Abs. 3 Nr. 1 PatG ist insofern für die semantische Information eines KNN rein deklaratorisch.[377]
Patentrechtlicher Schutz kommt daher für ein KNN nur anwendungsbezogen in Betracht, indem eine technische Lösung beansprucht wird – die Grundsätze zum Patentierbarkeit computerimplementierter Erfindung können entsprechend herangezogen werden.[378] Die Schutzmöglichkeit eines KNN durch das Patentrecht ist mithin eine Frage des Einzelfalls, wobei die neuen Prüfungsleit-

---

[375] *Loewenheim/Leistner*, in: Schricker/Loewenheim, UrhG, § 69a Rn. 14.

[376] *Loewenheim/Leistner*, in: Schricker/Loewenheim, UrhG, § 2 Rn. 47 f.; Springorum verneint allerdings bereits 1995 die urheberrechtliche Schutzfähigkeit der Gewichte eines KNN, da der Inhalt dem menschlichen Geist nicht zugänglich sei: „The content of the alleged presentation is not accessable [sic.] by the copyright law as it cannot be deduced from its form even though it is included in it without doubt.", *Springorum*, in: Brunnstein/Sint, Intellectual Property Rights and New Technologies. Proceedings of the KnowRight'95 Conference, S. 209.

[377] *Hauck/Cevc*, ZGE 2019, 135 (147).

[378] Siehe dazu nur *Hauck/Cevc*, ZGE 2019, 135 (148 ff.); *Nägerl/Neuburger/Steinbach*, GRUR 2019, 336 (339); *Ménière/Pihlajamaa*, GRUR 2019, 332 (334).

linien des Europäischen Patentamts eine verhältnismäßig patentfreundliche Praxis für KI-Systeme versprechen.[379]

Auf den ersten Blick könnte angenommen werden, dass Transparenzverpflichtungen für den Inhaber eines Patents an einem Erzeugnis oder Verfahren, das ein KNN beinhaltet, keine Gefahr darstellten. Denn das Patentrecht ist ja gerade getragen von dem Gedanken, dass Kenntnisse über Erfindungen im Tausch gegen das zeitlich begrenzte Schutzrecht preisgegeben werden.[380] Transparenz ist mithin für die Erlangung des Patents ohnehin erforderlich – es kommt jedoch auf deren Umfang an.

In einer Patentanmeldung muss die Erfindung gemäß § 34 Abs. 4 PatG und Art. 83 EPÜ „so deutlich und vollständig" offenbart werden, „dass ein Fachmann sie ausführen kann." Das Kriterium der Nacharbeitbarkeit bestimmt somit, wie viel der Anmelder von seinem KNN preisgeben muss. Die Anforderungen des Patentrechts an Offenbarung und Nacharbeitbarkeit sind daher auch für die Frage der Offenlegung des Geschäftsgeheimnisses am KNN von Interesse. Denn nach den soeben dargestellten Grundsätzen zur Erlangung des Geschäftsgeheimnisses fallen Offenbarung des Geschäftsgeheimnisses und Nacharbeitbarkeit im Sinne des Patentrechts zusammen.

Sowohl in der Rechtsprechung als auch in der Literatur sind Ausführungen zu konkreten Anforderungen, wie dem Erfordernis der Offenbarung bei Patentanmeldungen von KNN nachgekommen werden kann, bisher jedoch spärlich gesät.[381] Zumindest die Topologie dürfte zu offenbaren sein, die Beschwerdekammer des Europäischen Patentamts hat in einer Entscheidung sogar die Offenbarung von Trainingsdaten gefordert.[382] Bei computerimplementierten Erfindung ist nach dem BPatG die Offenbarung des Quellcodes allerdings für die Paten-

---

[379] *Europäisches Patentamt*, Richtlinien für die Prüfung im Europäischen Patentamt, Teil G-Kapitel II-1 ff.

[380] Siehe nur *Rogge/Melullis*, in: Benkard, Patentgesetz, Einl. Rn. 1.

[381] Siehe zur hinreichenden Offenbarung bei KNN EPA, 10.3.2000, T521/95 – *Pattern recognition/RDC JAPAN*; EPA, 12.5.2020, T 0161/18 – *Äquivalenter Aortendruck/ARC SEIBERSDORF*; zur Offenbarung bei KI allgemein *Hauck/Cevc*, ZGE 2019, 135 (154); *Ménière/Pihlajamaa*, GRUR 2019, 332 (335); *Nägerl/Neuburger/Steinbach*, GRUR 2019, 336 (339); *Tochtermann*, in: Kaulartz/Braegelmann, Rechtshandbuch Artificial Intelligence und Machine Learning, S. 327.

[382] zur Offenbarung der Trainingsdaten EPA, 12.5.2020, T 0161/18 – *Äquivalenter Aortendruck/ARC SEIBERSDORF*.

terteilung nicht erforderlich.[383] Die Entwicklung in diesem Bereich wird somit abzuwarten sein. Dass die Offenlegung der Gewichte eines trainierten KNN standardmäßig gefordert würde, zeichnet sich jedoch nicht ab.
Insofern dürften etwaige Transparenzpflichten selbst für den Inhaber eines Patents, das ein KNN beinhaltet, je nach Ausgestaltung eine Gefahr darstellen. Sofern die im Rahmen einer Transparenzpflicht offengelegte Information eines KNN patentrechtlichen Schutz genießt, stehen dem Patentinhaber die in den §§ 139 ff. PatG normierten Ansprüche zur Verfügung. Er kann mithin etwa auf Unterlassung der Benutzung oder auf Schadenersatz gegen einen Rechtsverletzer vorgehen. Aber auch wenn das Patentamt die Offenbarung lediglich der Topologie für die konkrete Anmeldung als ausreichend erachtet haben mag, so könnte eine Transparenzpflicht die Offenlegung weitergehender (geheimer) Informationen erforderlich machen.[384] Die Möglichkeiten des Schutzrechtsinhabers richten sich dann wiederum nach dem GeschGehG.

C. Darstellung der semantischen Information eines Künstlichen Neuronalen Netzes

Eine große Herausforderung im Bereich klassischer KI besteht darin, eine Beschreibung der physischen Welt zu entwerfen, die von intelligenten Systemen genutzt werden kann:

> „[C]onstructing a description of the physical world that can be used for intelligent system reasoning is one of the fundamental issues of Artificial Intelligence."[385]

Dies gelingt, indem implizites Wissen explizit gemacht wird durch symbolische Repräsentation. Dabei handelt es sich um die von den Anfängen der KI bis in die 1990er Jahre hinein gängige Methode, deren Ergebnis als *Symbolic Artificial Intelligence* oder auch *Good Old-Fashioned Artificial Intelligence* bezeichnet

[383] BPatG, Beschl. v. 8.7.2004, BPatGE 48, 238.

[384] Dass ein Erzeugnis sowohl geheime als auch patentgeschützte Aspekte umsetzt, ist nicht unüblich; vgl. *Kalbfus*, in: Harte-Bavendamm/Ohly/Kalbfus, GeschGehG, Einl. Rn. 180.

[385] *Flasiński*, Introduction to Artificial Intelligence, S. 21.

wird.[386] Diese Art von Modellbildung ist jedoch begrenzt und für viele Anwendungen von KI nicht zielführend. Verfahren des Maschinenlernen, darunter auch KNN, „lernen" daher durch Generalisierungen und speichern das „Gelernte" implizit.[387]

Die am Ende dieses Teils beschriebenen Methoden der Darstellung von in KNN vorhandener Information (XAI) folgen daher in gewisser Weise den gleichen Weg, den die GOFAI beschritt, nur in die entgegengesetzte Richtung: das implizite Wissen liegt nicht in der physischen Welt, sondern in der KI selbst vor[388] und soll explizit gemacht werden. Das heißt, die Regeln des KNN sollen herausgearbeitet werden. Adressat ist jedoch hier nicht die Maschine, sondern der Mensch, der die Wirkungsweise des KNN verstehen will.

Doch die unter dem Begriff XAI zusammengefassten Methoden bilden nur eine von vielen Möglichkeiten, wie ein KNN dargestellt werden kann, beziehungsweise von Formen, in denen es vorliegen kann.

Die verschiedenen Darstellungsmöglichkeiten lassen sich von der genauesten, konkreten Darstellung als Maschinencode bis zur ungenauesten, abstrakten Darstellung mit Hilfe der Techniken der XAI auf einer Skala anordnen. Dabei verläuft das Maß der faktischen semantischen Information, die der jeweiligen Darstellung entnommen werden kann, gegensätzlich zur Reproduzierbarkeit des Netzes. Während der Maschinencode die gesamte Information des KNN enthält, ist diese in ihm nicht explizit repräsentiert. Sie lässt sich nicht lokalisieren und die Funktionsweise des Netzes daher nicht ablesen. Die XAI-Methoden hingegen zeigen Teile der Information des KNN in expliziter Weise, bilden sie jedoch nur äußerst unvollständig ab.

Nach dem Grad der Abstraktion werden im Folgenden vier Darstellungsstufen eines KNN unterschieden:

Die *erste Darstellungsstufe* bildet der Maschinencode, der auch die Gewichte enthält.[389]

---

[386] Siehe zu den unterschiedlichen Methoden *Flasiński*, Introduction to Artificial Intelligence, S. 15 ff.; siehe zu Symbolic AI und GOFAI bereits oben, Kapitel 2 E.

[387] Vgl. zum Lernen durch Modellbildung und durch Generalisierung *Zech*, Weizenbaum Series 2020, 1 (30 f.).

[388] Flasiński spricht von „Models of knowledge representation formulated in an implicit way", *Flasiński*, Introduction to Artificial Intelligence, S. 225.

[389] Siehe zur Frage, ob die Gewichte im Maschinencode integriert oder separat gespeichert werden unten, Kapitel 5 A.

Auf der *zweiten Darstellungsstufe* stehen der Quellcode in einer Programmiersprache und die separat gespeicherten Gewichte.
Als *dritte Darstellungsstufe* kann dann die schon etwas abstraktere Beschreibung der Funktionsweise des Netzes anhand der Netzarchitektur, der verwendeten Funktionen und der Gewichte angesehen werden. Ansätze einer solchen Darstellung der Information dritter Stufe findet sich im ersten Teil dieser Arbeit.
Die *vierte Darstellungsstufe* ist die Art von Information, wie sie anhand der aktuell verfügbaren Möglichkeiten der *Explainable Artificial Intelligence* entsteht. Dabei handelt es sich also um diejenige Information, die für den Laien-Anwender einer KI verständlich sein soll und auf die die Transparenz-Debatte in Teilen abzielt.
Vor dem Hintergrund der Transparenz-Geheimnisschutz-Dichotomie stellen sich dann zwei Fragen: *ab* welchem Punkt wird die implizit repräsentierte Information explizit, wann kann sie also lokalisiert und für den menschlichen Betrachter verständlich repräsentiert werden? Diese Frage stellt sich aus dem Blickwinkel der Regulierung. Aus Sicht des Geheimnisschutzes hingegen muss gefragt werden: *bis zu* welchem Punkt ist in der Darstellung des KNN seine Information offenbar und mithin das Geschäftsgeheimnis durch Reproduzierbarkeit erlangt? Grundsätzlich liegt es in der Natur der Sache, dass sich Geheimnis und Transparenz ausschließen.[390] Die soeben formulierten Fragen zeigen jedoch, dass im Bereich Geheimnisschutz und Transparenz von Blackbox-Modellen eine differenzierte Betrachtung notwendig ist. Hier wird die bereits beschriebene Unterscheidung von Opazität aufgrund von Geheimhaltung, Opazität mangels Programmierkenntnissen und Opazität aufgrund der Diskrepanz zwischen algorithmischer Komplexität und menschlicher Auffassungsgabe relevant.[391]
Ziel der Transparenzforderungen hinsichtlich von KNN ist es zunächst immer, die Opazität durch Geheimhaltung zu überwinden. Je nach Adressaten der Transparenz müssen jedoch auch die anderen beiden Formen der Opazität überwunden werden. Während nämlich einem Experten für die Überprüfung der Funktionsweise eines KNN der Maschinencode ausreichen mag, muss für Transparenz gegenüber einem Laien die Opazität aufgrund von Komplexität beseitigt werden. Dabei handelt es sich bei der Transparenz gegenüber Laien um die für den Geheimnisschutz besonders relevante Art von Transparenz, da im

[390] Siehe dazu nur *Alexander*, MMR 2021, 690.
[391] Siehe dazu oben, Kapitel 3 A.

Gegensatz zur Transparenz gegenüber Behörden oder von diesen beauftragten Experten realistischerweise nur begrenzt auf Verschwiegenheitsverpflichtungen gesetzt werden kann. Gleichzeitig führen nicht beide Formen der Transparenz notwendigerweise zu einer Offenbarung des Geschäftsgeheimnisses, also der gesamten Information des KNN. Vielmehr werden die folgenden Kapitel zeigen, dass ein Mehr an Transparenz (als Gegenteil von Opazität durch Komplexität) ein Weniger an Geheimnisverlust (als Gegenteil von Opazität durch Geheimhaltung) bedeutet.

Kapitel 5

Erste Darstellungsstufe: Maschinencode

A. Technische Grundlagen und Darstellung der Information

Damit digitale Systeme Information verarbeiten können, muss sie in binär codierter Form vorliegen. Die Codierung erfolgt mithilfe eines sogenannten Compilers, der den in einer Programmiersprache geschriebenen Quellcode[392] in eine maschinenlesbare Form übersetzt.[393] Ergebnis ist der Maschinencode, der aus den Ziffern 0 und 1 besteht. Diese sind mögliche Zustände der kleinsten Informationseinheit, eines Bits.

Diese Kompilierung findet sowohl bei algorithmenbasierten Computerprogrammen als auch bei Künstlichen Neuronalen Netzen statt. Letztere weisen jedoch einige Besonderheiten auf, die sie von herkömmlichen Computerprogrammen unterscheiden. Bei der Programmierung eines KNN werden die Hyperparameter des Netzes, also Aufbau, Auswahl der Aktivierungsfunktionen etc.,[394] zunächst in einer Programmiersprache entworfen.[395] Ergebnis ist das untrainierte Netz, das in dieser Form gespeichert und auch vervielfältigt werden kann. Der Quellcode des untrainierten Netzes wird dann in Maschinencode kompiliert und erst dann wird das Netz mit Datensätzen trainiert. Die durch das Training berechneten Gewichte können dann entweder separat in einer Daten-

---

[392] Häufig auch etwas ungenau als „Programmcode" bezeichnet.

[393] *Schmidt*, in: Auer-Reinsdorff/Conrad, Handbuch IT- und Datenschutzrecht, § 1 Rn. 184.

[394] Häufig wird auch von der „Topologie" des Netzes gesprochen, so etwa *Hartmann/Prinz*, in: Taeger, Rechtsfragen digitaler Transformationen - Gestaltung digitaler Veränderungsprozesse durch Recht, S. 776 ff.

[395] Dafür wird häufig auf bereits bestehende Versatzstücke in Quellcode aus großen Datenbanken (sog. frameworks) wie etwa Tensorflow zurückgegriffen. Zum Verständnis des Prozesses soll hier dennoch der Ablauf beschrieben werden.

bank oder gleich im Maschinencode integriert gespeichert werden.[396] Letzteres ermöglicht die Speicherung und Vervielfältigung des trainierten Netzes in einer einzigen binären Datei.

Hier zeigt sich der Unterschied zu einem klassischen Computerprogramm: der Programmierer eines KNN greift zurück auf die ML-Technik, da er eine komplexe Aufgabe nicht durch einen Algorithmus lösen und mithin nicht präzise vorgeben kann, wie das Netz mit den Eingabedaten verfahren soll. Dadurch kann aber der von Hand entworfene Quellcode naturgemäß noch nicht die Spezifikationen – also die Gewichte – enthalten, die erst durch das Training festgelegt werden. Da zwischen dem Entwerfen des untrainierten Netzes in Quellcode und dem Training jedoch die Kompilierung liegt, schlägt sich das Trainingsergebnis im Allgemeinen nur im Maschinencode, nicht im Quellcode nieder.[397]

Gegenstand der Betrachtungen in diesem Kapitel ist der Maschinencode, der das trainierte Netz mitsamt den Gewichten enthält. In der juristischen Auseinandersetzung mit KNN werden zwar häufig die Topologie des untrainierten Netzes auf der einen und die Gewichte auf der anderen Seite betrachtet.[398] Allerdings wird in der Praxis meist das trainierte Netz inklusive der Gewichte in einer Binärdatei gespeichert.[399]

Der Maschinencode enthält dann notwendigerweise die vollständige semantische Information eines KNN. Sie ist binär codiert und kann als solche auch auf einem Bildschirm dargestellt und durch den Menschen wahrgenommen werden. Es handelt sich also um syntaktische Information, die – da der schier endlosen Aneinanderreihung der beiden Ziffern gemeinhin keine Bedeutung zu entnehmen ist – potenzielle semantische Information enthält. Potenziell ist die Information mangels expliziter Repräsentation, denn die Regeln der Informationsverarbeitung können nicht an einzelnen Symbolen des Maschinencodes abgelesen werden.

---

[396] *Ehinger/Stiemerling*, CR 2018, 761 (767).

[397] Siehe zum Ganzen *Hartmann/Prinz*, in: Taeger, Rechtsfragen digitaler Transformationen - Gestaltung digitaler Veränderungsprozesse durch Recht, S. 783, die von einer „Divergenz der Ausdrucksformen" sprechen.

[398] Siehe nur *Hartmann/Prinz*, in: Taeger, Rechtsfragen digitaler Transformationen - Gestaltung digitaler Veränderungsprozesse durch Recht, S. 776 ff.; *Kuß/Sassenberg*, in: Sassenberg/Faber, Rechtshandbuch Industrie 4.0 und Internet of Things: Praxisfragen und Perspektiven der digitalen Zukunft, S. 448 f.

[399] Dr. Florian Kaiser, persönliche Mitteilung.

B. Transparenzpflichten und Geheimnisschutz

Für den in großem Maße auf faktischer Geheimhaltung beruhenden Geheimnisschutz ist im Rahmen von Transparenzpflichten relevant, welche Information an wen und unter welchen Bedingungen preisgegeben werden muss. Dabei kommt es zudem maßgeblich darauf an, ob die herauszugebende Information die gesamte semantische Information des KNN enthält, anhand derer dann ein funktionsgleiches KNN reproduziert werden kann, oder ob nur eine Teileinformation herausgegeben werden muss.

Der Maschinencode enthält, wie oben untersucht, notwendigerweise die gesamte implizite Information des KNN.[400] Seine Offenlegung entspricht der Offenlegung der semantischen Information des Netzes und mithin der Erlangung des Geschäftsgeheimnisses. Dem steht nicht entgegen, dass die Information des KNN im Maschinencode nicht explizit und mithin für einen Menschen unverständlich wird. Denn die Information eines Geschäftsgeheimnisses kann auch potenzielle semantische Information sein.[401]

An die Offenlegung des Maschinencodes schließt sich die Möglichkeit der Reproduktion des KNN durch bloßes Kopieren der Binärdatei an. Dies führt zwar nach den bereits dargestellten Grundsätzen des Geschäftsgeheimnisrechts nicht automatisch zum Verlust des Geschäftsgeheimnisses. Doch das Risiko der Offenlegung und damit der Zugänglichkeit „ohne Weiteres" gemäß § 2 Nr. 1 lit. a GeschGehG steigt.

Aus den Auskunfts- und Informationsrechten in den Artikeln 13, 14 und 15 DSGVO ergibt sich kein Recht auf Übergabe des Maschinencodes. Ein solches widerspräche auch den beiden bereits dargestellten Parametern, die maßgeblich sind für die Bestimmung des Umfangs der über die involvierte Logik mitzuteilenden Information: der Schutz von Geschäftsgeheimnissen und die Verständlichkeit der Information (Art. 12 Abs. 1 DSGVO).

---

[400] Dies gilt zumindest dann, wenn (wie hier) von der Integration der Gewichte in den Maschinencode ausgegangen wird.

[401] Anders wohl teilweise die Bewertung durch US-Gerichte, wonach die Benutzer einer in Objektcode vorliegenden Software das Geschäftsgeheimnis an ihr nicht erlangen könnten, da der Objektcode für Menschen nicht verständlich sei. Siehe dazu *Surblyté*, in: Ullrich/Hilty/Lamping/Drexl, TRIPS plus 20: From Trade Rules to Market Principles, S. 738, unter Verweis auf Trandes Corp. v. Guy F. Atkinson Co., 798 F.Supp. 284, 288 (D. Md. 1992).

Weniger klar ist, welche Pflichten sich aus dem KI-VO-E für den Maschinencode ergeben. Wie gezeigt sieht der Entwurf umfangreiche Transparenzpflichten gegenüber Behörden und notifizierten Stellen, aber auch nicht unerhebliche Transparenzpflichten gegenüber Privaten vor. Jedenfalls gegenüber der ersten Gruppe kann auch die Herausgabe des Quellcodes verlangt werden. Da dieser jedoch nicht die wesentliche Information des KNN enthält, scheint es nicht ausgeschlossen, dass im Wege einer teleologischen Auslegung für eine Überprüfung des Netzes auch die Übergabe des Maschinencodes an die genannten öffentlichen Stellen gefordert würde. Eine Übergabe des Maschinencodes an Nutzer ist im KI-VO-E jedoch nicht vorgesehen.

Aufgrund der Reproduktionsmöglichkeit des KNN ist es auch sinnvoll, wenn eine Offenlegung des Maschinencodes allenfalls an Experten gefordert wird. Denn zum einen können nur Experten sinnvollerweise etwas mit dem Maschinencode eines KNN anfangen, etwa die Funktionsweise des Netzes überprüfen.[402] Zum anderen handelt es sich um das Herzstück des Geheimnisschutzes, dessen Herausgabe mit großen Risiken für den Geheimnisinhaber verbunden ist. Die Risiken des Geheimnisverlustes können zwar im Rahmen der Offenlegung an Behörden oder behördlich bestellte Experten durch Verschwiegenheitspflichten und Vertraulichkeitsvereinbarungen minimiert werden, was im KI-VO-E auch ausdrücklich verankert ist.[403] Denn die Herausgabe geheimer Information an eine Behörde führt nicht zu allgemeiner Bekanntheit im Sinne des GeschGehG, solange Dritte keine Rechte auf Einsicht durchsetzen können, was durch den Geheimnisschutz flankierende Regelungen wie § 29 Abs. 2 VwVfG und § 6 S. 2 IFG sichergestellt werden soll.[404]

Auch die Offenlegung des Maschinencodes an durch Vertrag oder Gesetz zur Verschwiegenheit Verpflichtete löst keine allgemeine Bekanntheit im Sinne des GeschGehG aus, solange nicht mit einer Weitergabe an Unbefugte zu rechnen ist.[405] Da es sich beim Geheimnisschutz jedoch nicht um ein absolutes Schutz-

[402] Häufig wird auch die Verfügbarkeit des Modells in Maschinencode nicht ausreichen, damit ein Experte die Funktionsweise des Modells vollumfänglich, etwa mit Blick auf Diskriminierungen, überprüfen kann. Siehe dazu auch unten im Vierten Teil.

[403] So explizit die Kommission: „Benötigen öffentliche und notifizierte Stellen für die Prüfung der Einhaltung der wesentlichen Pflichten Zugang zu vertraulichen Informationen oder zu Quellcodes, sind sie zur Wahrung der Vertraulichkeit verpflichtet", *Europäische Kommission*, Entwurf der KI-Verordnung, COM(2021) 206 final, S. 13.

[404] *Harte-Bavendamm*, in: Harte-Bavendamm/Ohly/Kalbfus, GeschGehG, § 2 Rn. 25.

[405] *Harte-Bavendamm*, in: Harte-Bavendamm/Ohly/Kalbfus, GeschGehG, § 2 Rn. 26.

recht handelt, können die Risiken nicht ganz ausgeschlossen werden. Die vertragswidrige, durch § 4 Abs. 2 Nr. 3 GeschGehG verbotene – und gegebenenfalls sogar strafbewährte (§ 23 GeschGehG, §§ 203, 204 StGB) – Offenlegung des Maschinencodes führt nicht per se zum Verlust des Geheimnisschutzes. Sobald die Information dadurch jedoch allgemein bekannt oder ohne weiteres zugänglich wird gem. § 2 Nr. 1a GeschGehG, tritt dieser dennoch ein.[406]
In Abhängigkeit von den durch den Empfänger des Maschinencodes getroffenen Sicherheitsvorkehrungen besteht auch das Risiko von Angriffen mit Techniken des Reverse Engineering, durch die gegebenenfalls ein funktionsgleiches Netz reproduzieren werden kann.[407] Auch das Reverse Engineering bewirkt dann für sich genommen nicht ohne Weiteres Offenkundigkeit, solange der Analyst das Ergebnis nicht offenlegt.[408] Und auch diese Offenlegung oder sogar Nutzung kann vertraglich verboten sein – § 3 Abs. 1 GeschGehG erlaubt nur das „Erlangen" – und gegen immaterialgüterrechtliche und lauterkeitsrechtliche Vorschriften verstoßen.[409] Allerdings hindert auch dies nicht unbedingt das faktische Element eine Offenlegung, durch die das Geschäftsgeheimnis verloren wäre.
Für die Herausgabe des Maschinencodes im Rahmen von Transparenzpflichten ergibt sich daraus Folgendes: erfolgt eine Herausgabe an eine Behörde, so unterliegen die betrauten Mitarbeiter Verschwiegenheitspflichten. Das Geschäftsgeheimnis bleibt daher grundsätzlich gewahrt.
Gleiches gilt bei einer Herausgabe an von der Behörde mit der Überprüfung der Funktionalität eines KNN betraute Dritte. Denn auch diese Dritte unterliegen öffentlich-rechtlichen oder vertraglichen Verschwiegenheitspflichten.
Eine Transparenzpflicht, die eine Herausgabe des Maschinencodes an Private vorsähe, wäre als mit dem Schutz von Geschäftsgeheimnissen unvereinbar abzulehnen. Denn grundsätzlich bildet der Geheimnisschutz eine Schranke für Transparenzpflichten. Selbst wenn davon im Einzelfall im Wege der Abwägung abgewichen werden kann, so wöge hier das Interesse des Informationssuchenden weit weniger schwer als das Interesse des Geheimnisinhabers. Denn das Risiko einer Offenlegung ist in dieser Fallkonstellation für den Geheimnisinhaber immens und können durch die an die Rechtsverletzung anknüpfenden Hand-

[406] Siehe zu der Frage, wann in diesem Fall von allgemeiner Bekanntheit auszugehen ist *Harte-Bavendamm*, in: Harte-Bavendamm/Ohly/Kalbfus, GeschGehG, § 2 Nr. 29.
[407] Siehe dazu unten, Kapitel 9.
[408] *Ohly*, in: Harte-Bavendamm/Ohly/Kalbfus, GeschGehG, § 3 Rn. 29.
[409] *Ohly*, in: Harte-Bavendamm/Ohly/Kalbfus, GeschGehG, § 3 Rn. 28, 31 ff.

lungsmöglichkeiten kaum abgemildert werden. Gleichzeitig dürfte der Nutzen einer Offenlegung des Maschinencodes für einen Privaten recht gering sein. Dies gilt zumindest für den Fall, dass in besonders sensiblen Anwendungsbereichen die Funktionalität des KNN im Rahmen eines behördlich etablierten Audit-Verfahrens anhand des Maschinencodes überprüft werden kann.

C. Urheberrechtlicher Schutz

Da der Maschinencode die vollständige semantische Information des KNN enthält, birgt seine Offenlegung das Risiko des faktischen Geheimnisverlustes. Es stellt sich daher die Frage nach Möglichkeiten außerhalb des Geheimnisschutzes, die es dem Geheimnisinhaber erlauben, seine Information vor unbefugter Verbreitung oder Nutzung zu schützen. Die Prüfung beschränkt sich hier auf das Urheberrecht, denn die patentrechtliche Schutzmöglichkeit bezieht sich auf die semantische Information und ist daher durch die bereits erfolgte Prüfung vollumfänglich abgedeckt.
Eine erhebliche Einschränkung der Schutzmöglichkeiten des Urheberrechts sei jedoch erneut betont: das Urheberrecht schützt grundsätzlich nur die konkrete Ausdrucksform und mithin syntaktische Information, der Inhalt bleibt frei.[410] Es kann daher den durch das Geschäftsgeheimnisrecht bestehenden Schutz der semantischen Information nur sehr eingeschränkt flankieren.
Voraussetzung für den Schutz als Computerprogramm gem. § 2 Abs. 1 Nr. 1 i. V. m. § 69a UrhG ist, dass es sich beim Maschinencode des trainierten KNN um ein Computerprogramm i. S. d. § 69a Abs. 1 UrhG handelt und dass er Ergebnis einer eigenen geistigen Schöpfung des Urhebers ist gem. § 69a Abs. 3 UrhG.
Für den Maschinencode klassischer Computerprogramme wird der Schutz als Computerprogramm unproblematisch bejaht.[411] ML-Techniken weisen jedoch einige Besonderheiten auf, die bei der Subsumtion berücksichtigt werden müssen.[412] Denn die Funktionalität des Programms wird nicht wie bei klassischer Software bereits durch den Programmierer abschließend festgelegt.

---

[410] *Zech*, Information als Schutzgegenstand, S. 246 f.

[411] EuGH, GRUR 2011, 220 (222) – *BSA/Kulturministerium*; zum alten Recht BGH, NJW 1986, 192 (196) – *Inkasso-Programm*; *Grützmacher*, in: Wandtke/Bullinger, UrhG, § 69a Rn. 11.

[412] Die Beurteilung der urheberrechtlichen Schutzfähigkeit von ML-Modellen ist höchst komplex und nicht abschließend geklärt. Im Folgenden sollen daher nur ein Überblick der Be-

Ganz allgemein kann der Umstand, dass ML-Modelle nicht mehr auf der klassischen Umsetzung eines Algorithmus durch einen Programmierer, sondern auf einem Training mit Daten beruhen, den Schutz durch das Urheberrecht nicht per se ausschließen. Die Schaffung einer „Methode zur Lösung" muss der Schaffung einer „Methode zur Findung einer Lösungsmethode" gleichwertig sein.[413] Diese Bewertung wird auch der Intention des Gesetzgebers gerecht, durch Verzicht auf eine Definition des Computerprogramms zukünftigen Entwicklungen im Bereich der Informatik Rechnung tragen zu können.[414]
Da das UrhG selbst keine Definition eines Computerprogramms enthält, wird überwiegend auf § 1 (i) der WIPO-Mustervorschriften für den Schutz von Computersoftware zurückgegriffen.[415] Danach ist ein Computerprogramm

> „eine Folge von Befehlen, die nach Aufnahme in einen maschinenlesbaren Träger fähig sind zu bewirken, dass eine Maschine mit informationsverarbeitenden Fähigkeiten eine bestimmte Funktion oder Aufgabe oder ein bestimmtes Ergebnis anzeigt, ausführt oder erzielt."

Wesentliches Merkmal eines Computerprogramms ist danach seine Steuerungsfunktion, das heißt es muss Steuerungsbefehle an einen Computer enthalten.[416] Der Maschinencode eines KNN enthält alle Befehle an den Computer, die zum Einsatz des KNN erforderlich sind. Dies gilt unproblematisch dann, wenn die Gewichte in den Code integriert und nicht separat gespeichert sind. Doch auch das KNN in binärem Code, in dem die Gewichte nicht enthalten sind, hat ausreichende Steuerfunktion um als Computerprogramm eingeordnet zu werden.[417] Das in Form von Maschinencode vorliegende trainierte KNN kann da-

---

sonderheiten gegenüber klassischen Computerprogrammen gegeben und die aktuelle Diskussion in der Wissenschaft grob skizziert werden. Einige Besonderheiten im Bezug auf KI analysieren auch *Hauck/Cevc*, ZGE 2019, 135 (159 ff.); *Hartmann/Prinz*, in: Taeger, Rechtsfragen digitaler Transformationen - Gestaltung digitaler Veränderungsprozesse durch Recht, S. 781 ff.

[413] *Hartmann/Prinz*, in: Taeger, Rechtsfragen digitaler Transformationen - Gestaltung digitaler Veränderungsprozesse durch Recht, S. 781.

[414] Siehe dazu nur *Wiebe*, in: Spindler/Schuster/Anton, Recht der elektronischen Medien, § 69a Rn. 3.

[415] So etwa BGH, NJW 1986, 192 (196) – *Inkasso-Programm*.

[416] *Loewenheim/Leistner*, in: Schricker/Loewenheim, UrhG, § 69a Rn. 2; OLG Hamburg, Urt. v. 12.3.1998, MMR 1999, 230 (231).

[417] So zum untrainierten Netz „in Computercode" *Ehinger/Stiemerling*, CR 2018, 761 (765).

mit wohl als Computerprogramm i. S. d. § 69a Abs. 1 UrhG angesehen werden.[418]

Die Besonderheit der ML-Technologie wird besonders relevant bei der Einschränkung des § 69a Abs. 3 UrhG, wonach das Computerprogramm als individuelles Werk das Ergebnis einer eigenen geistigen Schöpfung des Urhebers sein muss, um Schutz zu genießen.

Insgesamt ist die Frage, ab wann das trainierte Modell in seiner Gesamtheit ausreichend stark durch den Programmierer geprägt ist, um eine individuelle geistige Schöpfung und mithin als Computerprogramm urheberrechtlich schutzfähig zu sein, stark von der Art des Trainings abhängig und äußerst schwer zu beantworten.[419]

Eine geistige Schöpfung setzt voraus, dass das Werk „der Gedanken- und Gefühlswelt des Werkschaffenden entspringt".[420] Dies wird für den Maschinencode eines klassischen Computerprogramms grundsätzlich bejaht, auch wenn die Übersetzung des Quellcodes in den Maschinencode durch einen Compiler und nicht durch einen Menschen vorgenommen wird.[421] Denn trotz dieser Übersetzung entspringt der Maschinencode der Gedankenwelt des Programmierers, der seine Funktionen im Quellcode festgelegt hat.

Für ML-Technologien und die hier betrachteten KNN gilt dies jedoch nur bedingt: für die Gewichte des trainierten KNN wird das Merkmal der geistigen Schöpfung überwiegend abgelehnt.[422] Der Maschinencode dürfte jedoch auch ohne Ansehen der Gewichte grundsätzlich die erforderliche Schöpfungshöhe erreichen, sofern der Programmierer die ihm beim Entwurf des Quellcodes (als Grundlage des Maschinencodes) zur Verfügung stehenden Gestaltungsspielräume genutzt hat.[423]

---

[418] So wohl auch *Hartmann/Prinz*, in: Taeger, Rechtsfragen digitaler Transformationen - Gestaltung digitaler Veränderungsprozesse durch Recht, S. 783, zumindest für KNN, die nicht weiter lernen („offline"); die Frage ist „noch nicht geklärt" laut *Nägele/Apel*, in: Kaulartz/Braegelmann, Rechtshandbuch Artificial Intelligence und Machine Learning, S. 301.

[419] Allgemein zu dieser Schwierigkeit bei ML-Methoden *Drexl u. a.*, Max Planck Institute for Innovation and Competition Research Paper No. 21-10, S. 19; *Wiebe*, in: Spindler/Schuster/Anton, Recht der elektronischen Medien, § 69a Rn. 26.

[420] *Grützmacher*, in: Wandtke/Bullinger, UrhG, § 69a Rn. 34.

[421] *Grützmacher*, in: Wandtke/Bullinger, UrhG, § 69a Rn. 11.

[422] Siehe dazu ausführlich Kapitel 6, C II.

[423] So *Ehinger/Stiemerling*, CR 2018, 761 (766), die die Schöpfungshöhe auch beim Verwenden von Programmierframeworks nicht grundsätzlich ausschließen. Ebenso Hart-

Für den das trainierte Modell darstellenden Maschinencode ergibt sich daraus Folgendes: diejenigen Teile des Codes, welche die Gewichte enthalten, bleiben gemeinfrei. Sie können mithin aus dem Code kopiert und vervielfältigt werden. Denn die Vervielfältigung eines Werkteils ist nur unzulässig, sofern auch das Werkteil selbst urheberrechtlich geschützt ist. Eine differenzierte Betrachtung der einzelnen Werkteile ist mithin angebracht. Das verhindert jedoch nicht, dass der Maschinencode inklusive der integrierten Gewichte in seiner Gesamtheit urheberrechtlichen Schutz genießt und mithin auch in seiner Gesamtheit nicht vervielfältigt werden darf. [424]

Der Geheimnisinhaber könnte mithin zumindest hinsichtlich eines Teils seiner (offenbarten) Information auf die durch das Urheberrecht zur Verfügung gestellten Abwehrrechte hoffen. Doch diese Hoffnung wird durch die eigentliche Krux des urheberrechtlichen Schutzes enttäuscht: den Schutzumfang.

Der urheberrechtliche Grundsatz, dass lediglich die Form geschützt ist, bedeutet für den Bereich der Computerprogramme, dass ihre Funktionalität aus dem Schutzumfang herausfällt.[425] Denn § 69a Abs. 2 S. 2 UrhG nimmt „Ideen und Grundsätze, die einem Element eines Computerprogramms zugrunde liegen," ausdrücklich vom Schutzumfang aus.

Der Maschinencode fällt mithin in seiner konkreten Ausgestaltung als Übersetzung aus dem in einer bestimmten Programmiersprache geschriebenen Quellcode in den Schutzbereich des § 69a UrhG, die ihm zugrundeliegenden „Ideen und Grundsätze" jedoch nicht. Das bedeutet konkret, dass die individuelle Organisation und Struktur des Maschinencodes, also die konkrete Folge von Steuerungsbefehlen in der jeweiligen Programmiersprache geschützt sind.[426] Im Code enthaltene Algorithmen fallen grundsätzlich aus dem Schutzbereich heraus, kön-

---

mann/Prinz, die für den Entwurf der Topologie sogar einen größeren Gestaltungsspielraum annehmen als für die Programmierung eines klassischen Computerprogramms, *Hartmann/Prinz*, in: Taeger, Rechtsfragen digitaler Transformationen - Gestaltung digitaler Veränderungsprozesse durch Recht, S. 782.

[424] Springorum stellt bereits 1995 fest: „It can be stated that it does not lead to a satisfactory result to view the neural networks as a whole, from the point of view of the ability for copyright protection.", *Springorum*, in: Intellectual Property Rights and New Technologies. Proceedings of the KnowRight'95 Conference, S. 210; siehe zum Ganzen die differenzierte Analyse bei *Ehinger/Stiemerling*, CR 2018, 761 (767); ebenso wohl auch *Apel/Kaulartz*, RDi 2020, 24 (27).

[425] Apel/Kaulartz sprechen davon, dass „der eigentliche Witz" von Modellen „- ihre technische Funktionalität" nicht geschützt ist, *Apel/Kaulartz*, RDi 2020, 24 (28).

[426] Siehe zu Computerprogrammen im Allgemeinen nur *Spindler*, in: Schricker/Loewenheim, UrhG, § 69a Rn. 10; speziell zu neuronalen Netzen *Ehinger/Stiemerling*, CR 2018, 761 (765).

nen jedoch „in der Art und Weise der Implementierung und Zuordnung zueinander" urheberrechtlichen Schutz genießen.[427]

Für künstliche neuronale Netze bedeutet das: der Entwickler kann sich gegen die Vervielfältigung des konkreten Maschinencodes sowie dessen individueller Organisation und Struktur wehren.[428] Die Idee hinter der konkreten Ausgestaltung des KNN, also die Anzahl der Schichten und der Neuronen pro Schicht, die Struktur der Schichten, die Neuronen selbst, die verwendeten Funktionen sowie – mangels Schöpfungshöhe – die in den Maschinencode integrierten Gewichte bleiben frei.[429]

Der Schutz des Maschinencodes eines trainierten KNN durch das Urheberrecht ist mithin lückenhaft.[430] Die Abgrenzung von geschützter konkreter Ausgestaltung und ungeschützten Ideen ist zudem, wie bei klassischen Computerprogrammen auch, schwierig und eine Frage des Einzelfalls.[431]

---

[427] Sog. „Gewebetheorie", BGH, Urteil v. 4.10.1991, GRUR, 449 (453) – *Betriebssystem*.

[428] So auch Ehinger/Stiemerling zum Programmcode eines KNN *Ehinger/Stiemerling*, CR 2018, 761 (765).

[429] *Ehinger/Stiemerling*, CR 2018, 761 (766); *Grützmacher*, in: Wandtke/Bullinger, UrhG, § 69a Rn. 21.

[430] So das weit überwiegende Fazit in der Wissenschaft, vgl. nur *Apel/Kaulartz*, RDi 2020, 24 (29).

[431] *Ehinger/Stiemerling*, CR 2018, 761 (765).

Kapitel 6

Zweite Darstellungsstufe: Quellcode und Gewichte

Der Entwurf des Quellcodes ist im Prozess der Programmierung eines KNN der Erstellung des Maschinencodes zeitlich vorgelagert, der erst durch die Kompilierung des Quellcodes entsteht.[432] Die Darstellung erfolgt hier mithin nicht chronologisch, sondern richtet sich nach dem Maß, in dem in einer Darstellungsform eines KNN die semantische Information des Netzes enthalten ist, also nach der Abstraktionsstufe.
Neben dem Quellcode sollen auf der zweiten Darstellungsstufe auch die Gewichte betrachtet werden. Im vorangegangenen Teil wurden sie zwar als Teil des Maschinencodes analysiert. Allerdings besteht auch die Möglichkeit, sie separat als Datei zu speichern, auf die während des Programmablaufs dann zugegriffen wird. Auch für diesen Fall sollen ihre Schutzmöglichkeiten untersucht werden.[433]

A. Technische Grundlagen und Darstellung der Information

I. Quellcode

Der Quellcode ist der in einer beliebigen Programmiersprache verfasste, von Menschen lesbare Text eines Computerprogramms.[434] Dies gilt auch für den Quellcode eines KNN, der grundsätzlich keine Besonderheiten gegenüber dem

[432] Siehe dazu oben, Kapitel 5.

[433] Mit Blick auf ihren Abstraktionsgrad könnten die separat gespeicherten Gewichte auch auf der ersten Darstellungsstufe dargestellt werden. Die Einordnung erfolgt hier mit Blick auf dem Geheimnisschutz jedoch auf eine Stufe mit dem Quellcode, da beide Informationen zusammengenommen eine Reproduktion des KNN ermöglichen (dazu unten).

[434] *Schmidt*, in: Auer-Reinsdorff/Conrad, Handbuch IT- und Datenschutzrecht, § 1 Rn. 148.

Quellcode eines Computerprogramms aufweist, das keine ML-Methode verwendet. Auf eine Definition von Datenstrukturen folgt die Spezifikation von Verarbeitungsvorschriften und die Definition von Konstanten.[435]
Es wurde bereits erläutert, dass der Quellcode eines KNN dessen semantische Information nur in Teilen enthält. Aus ihm lassen sich die Hyperparameter des Netzes herauslesen: Anzahl der Schichten und Neuronen, Verbindungsstruktur, verwendete Aktivierungs- und Propagierungsfunktion und Lernalgorithmus. Was jedoch nicht im Quellcode enthalten ist, sind die Gewichte der einzelnen Verbindungen, also das Ergebnis des Trainings mit einem Trainingsdatensatz.[436] Der Quellcode kann daher nur Ausdruck des untrainierten, nicht des trainierten Modells sein. Gemeinsam mit den separat gespeicherten Gewichten kann er jedoch das trainierte Netz abbilden.
Der Quellcode enthält gleichzeitig (noch) keine explizite Repräsentation des im KNN gespeicherten Wissens. Seine Funktionalität kann nicht durch einen Blick auf einen Steuerungscode überprüft werden.[437]

II. Gewichte (Datei)

Bei den Gewichten handelt es sich um Zahlenwerte, die nach Abschluss des Trainings als Datei gespeichert und vervielfältigt werden können.[438] Die Speicherung kann als binäre Datei erfolgen, die dann durch einen Menschen nicht lesbar ist. Sie kann jedoch auch in einem Dateiformat gespeichert werden, das dann die Zahlenwerte für Menschen lesbar anzeigt. Die Darstellung erfolgt dann häufig in Form einer Tabelle, aus der sich die Gewichte (und Schwellenwerte) der einzelnen Neuronen jeder Schicht ablesen lassen.[439]
Die Gewichte enthalten einen Großteil der Information des KNN. Denn aus ihnen lassen sich nicht nur die Gewichtungen aller Verbindungen im Netz, sondern dadurch naturgemäß auch alle Information über die Architektur des Netzes ablesen. Lediglich die Berechnung innerhalb der einzelnen Neuronen geht aus ihnen nicht hervor. Diese erfolgt jedoch üblicherweise anhand einer kleinen

[435] *Nebel/Stiemerling*, CR 2016, 61 (62).
[436] Siehe zu dieser Diskrepanz zwischen Quellcode und Maschinencode bei KNN eingehend oben, Kapitel 5 A.
[437] Siehe zu dieser Folge impliziter Wissensrepräsentation in ML-Modellen *Zech*, Weizenbaum Series 2020, 1 (33 f.).
[438] Siehe dazu *Ehinger/Stiemerling*, CR 2018, 761 (764).
[439] Siehe zu dieser Darstellungsform das folgende Kapitel.

Zahl gängiger Funktionen und stellt mithin keine Hürde für einen Nachbau des Netzes dar.[440]
Die semantische Information des KNN ist in den Gewichten mithin numerisch gespeichert, sie ist jedoch weiterhin nicht explizit repräsentiert.

B. Transparenzpflichten und Geheimnisschutz

Aus dem Quellcode lassen sich die Architektur des Netzes, die verwendeten Funktionen und der Schwellenwert ablesen. Er enthält mithin bereits sensible Information, die einen Teil des Geschäftsgeheimnisses am KNN ausmachen.
Allein anhand der aus dem Quellcode ablesbaren Hyperparameter des Netzes lässt sich jedoch nicht ohne Weiteres ein äquivalentes KNN reproduzieren. Seine Erlangung führt mithin nicht zur Erlangung des Geschäftsgeheimnisses am KNN, das hier anhand der Reproduzierbarkeit gemessen wird.
Etwas anderes gilt für die Gewichte: Aus der Datei, in der die Gewichte gespeichert sind, lässt sich für jedes Neuron ablesen, wie die Eingaben, die es erreichen, gewichtet werden. Daher kann aus den Gewichtungen naturgemäß auch die Anzahl der Neuronen und Schichten abgelesen werden. Mit diesem Wissen ließe sich wohl ein funktionsgleiches Netz nachbauen, wenn auch mit einigem Aufwand. Die fehlenden Informationen über Aktivierungsfunktion und Schwellenwert können durch Tests ermittelt werden.
Noch riskanter ist für den Geheimnisinhaber die Erlangung sowohl des Quellcodes als auch der Gewichte durch Dritte. Denn Quellcode und Gewichte machen das KNN aus, beinhalten sein gesamtes „Wissen“ und ermöglichen seine Nacharbeitung. Eine Einschränkung kann lediglich die Unkenntnis über die erforderliche Datenvorbereitung mit sich bringen. Zudem können für einen erfolgversprechenden Einsatz Informationen über Trainingsdaten notwendig sein, da das trainierte Netz möglicherweise für den gewollten Einsatzbereich sonst nicht valide ist.[441]

---

[440] Dies gilt zumindest, wenn in allen Neuronen die gleiche Aktivierungsfunktion benutzt wurde. Andernfalls ergibt sich eine exponentielle Kombinationsmöglichkeit, welche einen Nachbau praktisch unmöglich machen dürfte.

[441] So müssen beispielsweise Scoring-Daten regional angepasst werden, *Gesellschaft für Informatik*, Technische und rechtliche Betrachtungen algorithmischer Entscheidungsverfahren, S. 51.

Die zwei untersuchten Regulierungsregime verhalten sich in Bezug auf den Quellcode unterschiedlich: während im Rahmen der Informations- und Auskunftsrechte der DSGVO eine Offenlegung des Quellcodes überwiegend als zu weitgehend abgelehnt wird, sieht der KI-VO-E eine Herausgabe des Quellcodes explizit vor, wenn auch nur gegenüber der Marktüberwachungsbehörde und, auf Antrag, gegenüber notifizierten Stellen.
Die Herausgabe des Quellcodes allein wäre für den Geheimnisinhaber verhältnismäßig unproblematisch. Denn ohne die Parametrierung durch die Gewichte lässt sich aus ihm als bloße Struktur des Netzes nicht dessen Funktionsweise ableiten. Daher lässt sich das Netz jedoch auch nicht auf etwaige Fehler untersuchen – die Herausgabe des Quellcodes als dem Anschein nach „schärfstes Schwert" des KI-VO-E ergibt daher für KNN nur in Verbindung mit der Herausgabe der Gewichte Sinn.
Auch wenn die Herausgabe der Gewichte im KI-VO-E nicht eindeutig geregelt ist, dürfte daher jedenfalls gegenüber den Marktüberwachungsbehörden eine Übergabe auch der Gewichte erforderlich sein.[442] Ob dies auch gegenüber notifizierten Stellen und anderen zuständigen Behörden und öffentlichen Stellen gilt, ist schon fraglicher. Dem Inhalt der technischen Dokumentation (Anhang IV KI-VO-E) lässt sich dazu keine eindeutige Aussage entnehmen. Dort ist lediglich vorgesehen, dass offengelegt werde muss „was das System optimieren soll und welche Bedeutung den verschiedenen Parametern dabei zukommt" (Nr. 2b). Die Bedeutung der verschiedenen Parameter könnte möglicherweise auch durch Angabe der näherungsweisen Gewichte nachgekommen werden, ohne dass Zugriff auf die entsprechende Datei gewährt würde. Die Übergabe dieser Datei selbst dürfte jedoch aus bereits genannten Gründen im Rahmen des auf begründeten Antrag zu gewährenden Zugangs zum Quellcode erforderlich sein (vgl. Anhang VII 4.5.).
Der im KI-VO-E vorgesehene gestaffelte Zugang zum Quellcode – Marktüberwachungsbehörde auf begründetes Verlangen (Art. 64 Abs. 2 KI-VO-E), notifizierte Stelle auf begründeten Antrag (Anhang VII 4.5. KI-VO-E), andere Behörden und öffentliche Stellen nicht – deutet darauf hin, dass zumindest gegenüber der letztgenannten Gruppe ein Zugang zu den Gewichten eines KNN nicht zu gewähren ist. Denn diese machen die „Kronjuwelen" des KNN aus und dürften daher ebenso restriktiv gehandhabt werden, wie es der KI-VO-E für den

[442] Auch Hacker spricht sich für eine Dokumentationspflicht der Gewichte für Auditierungszwecke aus, *Hacker*, NJW 2020, 2142 (2143).

Quellcode vorsieht. Es ließe sich mithin durchaus argumentieren, dass die technische Dokumentation, die nicht nur den notifizierten Stellen, sondern über Art. 23 KI-VO-E wohl auch zuständigen nationalen Behörden zur Verfügung gestellt werden muss, nicht die Gewichte enthalten muss. Auch den anderen öffentlichen Stellen wäre dann im Rahmen ihrer Prüfpflichten kein Zugang zu den Gewichten zu gewähren.

Über die Transparenzvorschrift des Art. 13 KI-VO-E scheint es zwar nicht ausgeschlossen, dass die Gewichte sogar gegenüber dem Nutzer eines KI-Systems herausgegeben werden müssten. Den Nutzern sollen in der Gebrauchsanweisung des KI-Systems unter anderem dessen Merkmale zur Verfügung gestellt werden. Die Information soll „präzise, vollständig[...], korrekt[...] und eindeutig[...]" sein und in einer „verständlichen Form" vorliegen gemäß Art. 13 Abs. 2 KI-VO-E. Wie oben dargelegt, dürfte die Einschränkung der Verständlichkeit jedoch aufgrund des anderen Empfängerhorizonts im Vergleich zur DSGVO hier weniger beschränkend wirken.

Die Merkmale, die notwendig sind, „damit die Nutzer das System angemessen verstehen und verwenden können" gemäß Art. 13 Abs. 1 KI-VO-E, könnten daher auch die Gewichte umfassen. Ob aus der gegenüber Nutzern vorgesehenen Transparenz- und Informationspflicht tatsächlich eine derart weitgehende Offenlegung gegenüber den Nutzern von KI-Systemen folgen wird, werden erst die abschließenden Verhandlungen der Norm zeigen.[443] Sofern es nicht zu einer weiteren Präzisierung durch den Normgeber kommt, wird darüber hinaus einschlägige Rechtsprechung abzuwarten sein. Zum jetzigen Zeitpunkt erscheint es jedenfalls nicht ausgeschlossen, dass umfangreiche Informationen auch über die Gewichte eines KNN zur Verfügung gestellt werden müssen. Dies dürfte jedoch weniger in der in diesem Kapitel behandelten Form als Binärdatei, sondern vielmehr anhand der im nächsten Kapitel beschriebenen Formen erfolgen.

Mit der obigen Argumentation betreffend eine Herausgabe der Gewichte an Behörden und öffentliche Stellen dürfte sich eine Herausgabe an die Nutzer jedoch erst recht verneinen lassen.

Denn diese wäre für den Geheimnisinhaber äußerst riskant, auch wenn alle Informationsempfänger im KI-VO-E zur Verschwiegenheit verpflichtet sind. Grund ist, dass die Möglichkeit des faktischen Geheimnisverlusts fortbesteht. Dies gilt umso mehr, als Nutzer im Sinne des KI-VO-E mittlerweile auch natür-

[443] Aktuell wird der KI-VO-E noch im Europäischen Parlament verhandelt, Gegenstand der vorliegenden Untersuchung ist die Allgemeine Ausrichtung des Rates vom 6.12.2022.

liche Personen sind, die ein KI-System privat nutzen. Dagegen ist das Risiko, das der Geheimnisinhaber mit einer Herausgabe an die zuständigen Behörden, notifizierten Stellen und andere öffentliche Stellen eingeht, zwar geringer, jedoch ebenfalls nicht zu vernachlässigen.

C. Urheberrechtlicher Schutz

I. Quellcode

Mit dem Quellcode wird nur ein Teil der als Geschäftsgeheimnis geschützten Information des KNN erlangt. Die Aussage, der „Quellcode einer Internet-Anwendung [sei] ein „Kronjuwel", wenn das Geschäftsmodell des Unternehmens auf einer spezifischen Funktion der Software basiert"[444], gilt daher für KNN nur sehr eingeschränkt. Dennoch kann seine Herausgabe die Erlangung des Geschäftsgeheimnisses als Ganzes begünstigen, sodass auch seine Schutzmöglichkeiten für die vorliegende Untersuchung von Interesse sind.
Für den urheberrechtlichen Schutz des Quellcodes eines KNN gelten grundsätzlich die bereits für den Maschinencode dargestellten Maßstäbe, da der Maschinencode durch Kompilierung des Quellcodes entsteht. Dass die Funktionalität von Maschinencode und Quellcode bei ML-Verfahren auseinanderfällt, spielt hier keine Rolle. Denn die Gewichte, die die Funktionalität maßgeblich bestimmen, fallen nach hiesiger Auffassung ohnehin nicht in den Schutzbereich des § 69a UrhG.[445] Der Quellcode hat mithin dieselbe Funktionalität wie der von Gewichten losgelöst betrachtete Maschinencode und kann mit Blick auf die Voraussetzungen des § 69a UrhG wie dieser betrachtet werden. Er ist Computerprogramm im Sinne des § 69a Abs. 1 UrhG und weist üblicherweise die erforderliche Schöpfungshöhe im Sinne des § 69a Abs. 3 UrhG auf.[446]
Gemäß § 69a Abs. 1 S. 2 UrhG ist auch hier nicht die Idee der Struktur, sondern nur die konkrete Ausdrucksform geschützt, das heißt der Quellcode in der kon-

[444] *Maaßen*, GRUR 2019, 352 (356).
[445] Dazu sogleich.
[446] *Ehinger/Stiemerling*, CR 2018, 761 (765); *Söbbing*, CR 2020, 223 (228); *Kuß/Sassenberg*, in: Sassenberg/Faber, Rechtshandbuch Industrie 4.0 und Internet of Things: Praxisfragen und Perspektiven der digitalen Zukunft, S. 448; die Schutzfähigkeit des Quellcodes eines KNN gemäß § 69a UrhG wird gemeinhin bejaht, *Söbbing*, MMR 2021, 111 (114); siehe zur Individualität von Quellcode die ausführliche Analyse bei *Hoeren/Wehkamp*, CR 2018, 1 (1 ff.).

kreten Programmiersprache sowie die konkrete Organisation und innere Struktur des Programms. Eine „funktionelle Imitation“[447] des KNN anhand des Quellcodes stellt mithin keine zustimmungsbedürftige Handlung gemäß § 69c UrhG dar.

II. Gewichte

Die Gewichte eines trainierten KNN gehen nicht auf eine eigene geistige Leistung des Entwicklers zurück, sondern werden während des Trainings durch den Computer berechnet. Sie beruhen damit zwar auf der Topologie des Netzes und der Datenauswahl und -aufbereitung des Entwicklers, sind aber streng genommen nicht Produkt eines menschlichen Geistes.
Ihre Schutzfähigkeit als Computerprogramm nach § 69a UrhG wird daher mangels geistiger Schöpfung ganz überwiegend abgelehnt.[448] Zwar kann grundsätzlich auch das, was mit Werkzeugen hergestellt wird, noch zugerechnet werden - es muss aber noch „durch die menschliche Leistung geprägt“ sein.[449]
Teilweise wird angenommen, dass die alles andere als banale Datenauswahl und -aufbereitung ausreichend ist für die Bejahung einer geistigen Schöpfung im Sinne des § 69a Abs. 3 UrhG und dass somit das gesamte trainierte Netz als Computerprogramm schutzfähig ist.[450] Andere unterscheiden nach der Art des Trainings, also nach überwachtem und bestärkendem Lernen auf der einen und unüberwachtem Lernen auf der anderen Seite, und verneinen nur für trainierte Netze, die auf unüberwachtem Lernen beruhen, einen urheberrechtlicher Schutz.[451] Eine persönliche Schöpfung wird vereinzelt auch dann noch ange-

---

[447] *Grützmacher*, in: Wandtke/Bullinger, UrhG, § 69c Rn. 11.

[448] So allgemein zu Lernergebnissen bzw. automatisiert erzeugten Elementen *Bischof/Intveen*, ITRB 2019, 134 (136); *Apel/Kaulartz*, RDi 2020, 24 (27); *Grützmacher*, in: Wandtke/Bullinger, UrhG, § 69a Rn. 21; zur Frage, ob nicht schon mangels Steuerungsbefehlen die Eigenschaft als Computerprogramm verneint werden müsste, siehe die ausführliche Analyse bei *Ehinger/Stiemerling*, CR 2018, 761 (766 f.).

[449] *Kuß/Sassenberg*, in: Sassenberg/Faber, Rechtshandbuch Industrie 4.0 und Internet of Things: Praxisfragen und Perspektiven der digitalen Zukunft, S. 448.

[450] *Hartmann/Prinz*, in: Taeger, Rechtsfragen digitaler Transformationen - Gestaltung digitaler Veränderungsprozesse durch Recht, S. 785.

[451] *Kuß/Sassenberg*, in: Sassenberg/Faber, Rechtshandbuch Industrie 4.0 und Internet of Things: Praxisfragen und Perspektiven der digitalen Zukunft, S. 448.

nommen, wenn die Auswahl der Parameter noch durch den Programmierer und lediglich die Gewichtung im Wege des Trainings erfolgt.[452]
Der Schutz der Gewichte als Computerprogramm nach § 69a UrhG ist mithin noch nicht abschließend geklärt, wird jedoch eher zu verneinen sein.
In Betracht kommen allerdings noch andere Möglichkeiten des urheberrechtlichen Schutzes, namentlich der Schutz als Datenbankwerk gemäß § 4 Abs. 2 UrhG oder als Datenbank gemäß § 87 UrhG.
Der Schutz als Datenbankwerk gemäß § 4 Abs. 2 UrhG setzt jedoch, da es sich um einen Unterfall des Sammelwerks gemäß § 4 Abs. 1 UrhG handelt, eine persönliche geistige Schöpfung voraus.[453] Ein Schutz der Gewichte gemäß § 4 Abs. 2 UrhG scheitert daher aus den bereits bei der Prüfung der Schutzfähigkeit nach § 69a UrhG angeführten Gründen.[454]
Anders könnte die Bewertung für das Leistungsschutzrecht des Datenbankherstellers gemäß § 87a UrhG ausfallen, das gerade keine persönliche geistige Schöpfung voraussetzt, da es allein dem Investitionsschutz dient.[455]
Allerdings wird die Schutzfähigkeit der Gewichte über § 87a UrhG aus anderen Gründen verneint. Vereinzelt wird bereits die gemäß § 87a Abs. 1 S. 1 UrhG erforderliche Einzelzugänglichkeit für nicht gegeben erachtet.[456] Zumeist wird der Schutz jedoch mangels Unabhängigkeit der einzelnen Gewichte abgelehnt.[457] Denn das Vorliegen unabhängiger Elemente gemäß § 87a Abs. 1 UrhG würde voraussetzen, dass die einzelnen Gewichte voneinander getrennt werden könn-

---

[452] So wohl *Loewenheim/Leistner*, in: Schricker/Loewenheim, UrhG, § 69a Rn. 15.

[453] *Leistner*, in: Schricker/Loewenheim, UrhG, § 4 Rn. 50; *Wiebe*, in: Spindler/Schuster/Anton, Recht der elektronischen Medien, § 4 Rn. 14 ff.

[454] Für eine ausführliche Analyse der weiteren Tatbestandsmerkmale des § 4 UrhG siehe *Ehinger/Stiemerling*, CR 2018, 761 (768 f.).

[455] *Dreier*, in: Dreier/Schulze: UrhG, Vor § 87a Rn. 1.

[456] So bereits *Hornung*, Die EU-Datenbank-Richtlinie und ihre Umsetzung in das deutsche Recht, S. 75; wohl auch *Grützmacher*, Urheber-, Leistungs- und Sui-generis-Schutz von Datenbanken, S. 65 f.; *Vogel*, in: Schricker/Loewenheim, UrhG, § 87a Rn. 13.

[457] *Ehinger/Stiemerling*, CR 2018, 761 (769); *Hartmann/Prinz*, in: Taeger, Rechtsfragen digitaler Transformationen - Gestaltung digitaler Veränderungsprozesse durch Recht, S. 786; *Hauck/Cevc*, ZGE 2019, 135 (162) m. w. N. zur Verneinung der Unabhängigkeit bei mathematischen Modellen im Allgemeinen; *Nägele/Apel*, in: Kaulartz/Braegelmann, Rechtshandbuch Artificial Intelligence und Machine Learning, S. 310; nach Kuß/Sassenberg eine Frage des Einzelfalls: *Kuß/Sassenberg*, in: Sassenberg/Faber, Rechtshandbuch Industrie 4.0 und Internet of Things: Praxisfragen und Perspektiven der digitalen Zukunft, S. 449.

ten, ohne dass dadurch der Wert ihres Inhalts geschmälert würde.[458] Der Wert der Gewichte ergibt sich jedoch gerade aus ihrer einzigartigen Kombination, die die Funktionsfähigkeit des KNN gewährleistet. Die Gewichtung einer einzelnen Verbindung zwischen zwei Neuronen ist daher wertlos.[459]

[458] Siehe zum Erfordernis der Unabhängigkeit nur *Dreier*, in: Dreier/Schulze: UrhG, § 87a Rn. 6.

[459] Siehe dazu nur *Ehinger/Stiemerling*, CR 2018, 761 (768 f.).

Kapitel 7

Dritte Darstellungsstufe: Beschreibung, Graph, Formeln

Die dritte Darstellungsstufe hebt die Information eines KNN auf ein Abstraktionsniveau, das für die Erfüllung von Transparenzpflichten gegenüber Laien auf den ersten Blick geeigneter erscheint als die Darstellungen der ersten beiden Stufen.
Bei den unter die dritte Stufe gefassten Darstellungsmöglichkeiten handelt es sich um Beschreibungen des KNN. Anders als die auf den ersten beiden Stufen dargestellten Codes sind es mithin nicht Teile des Computerprogramms oder Daten, mit dem das KNN ausgeführt wird. Vielmehr handelt es sich um Darstellungsweisen, wie sie teilweise auch von Programmierern im Zuge des Entwurfs eines KNN genutzt werden.[460]

A. Darstellung der Information

I. Beschreibung mittels natürlicher Sprache[461]

Viele Bestandteile eines künstlichen neuronalen Netzes können in natürlicher Sprache beschrieben werden.[462] So können die Anzahl der Schichten und der Neuronen je Schicht sowie die Art der Informationsverarbeitung (vorwärtsge-

[460] Außer Acht gelassen wird hier die Darstellung in Pseudocode, da eine auf diesbezügliche Transparenzpflicht wenig Sinn ergäbe. Für Laien ist der Pseudocode zwar verständlicher als der Quellcode, auch er bietet jedoch keine nachvollziehbare Erklärung der Informationsverarbeitung im Netz. Für Experten hingegen ist der Pseudocode üblicherweise weniger präzise als der Quellcode und eignet sich nicht zu einer automatischen Überprüfung des Netzes.

[461] Der Begriff wird hier in Abgrenzung zu Programmiersprache verwendet.

[462] So zur Topologie bereits *Springorum*, in: Brunnstein/Sint, Intellectual Property Rights and New Technologies. Proceedings of the KnowRight'95 Conference, S. 209.

richtet oder rückgekoppelt) einfach sprachlich ausgedrückt werden.[463] Auch die verwendeten Formeln können schlicht benannt („Heavisidesche Stufenfunktion als Aktivierungsfunktion") oder aber mathematisch ausgedrückt werden.[464] Diese Art der Beschreibung ist auch in Patentanmeldungen üblich.[465]
Schwieriger gestaltet sich jedoch die Beschreibung der Verbindungen zwischen den Neuronen, sofern es sich nicht um ein sehr kleines oder aber ein einfaches vorwärtsgerichtetes Modell handelt, bei dem beispielsweise alle Neuronen einer Schicht mit allen Neuronen der nächsten Schicht verbunden sind, jedoch nicht die Neuronen einer Schicht untereinander. Unübersichtlich würde auch eine Beschreibung der Gewichte der einzelnen Verbindungen in natürlicher Sprache, weshalb hierfür andere Methoden vorzugswürdig sind (dazu sogleich).
Mittels natürlicher Sprache kann jedoch ein Großteil der Hyperparameter des Netzes bereits präzise beschrieben und ein wichtiger Teil der Netzinformation offengelegt werden. Die vollständige semantische Information des Netzes lässt sich auf diese Weise allerdings kaum offenlegen, da sie die genaue Beschreibung der gewichteten Verbindungen erfordern würden.
Die Information ist auch in der Beschreibung des Netzes weiterhin nicht explizit repräsentiert. Denn auch wenn durch diese Art der Darstellung der Aufbau des Netzes und seine Funktionsweise nachvollzogen werden kann, so gibt sie keinen Einblick in die Bedeutung der Informationsverarbeitung im Netz.

II. Darstellung als Graph

Künstliche neuronale Netze sind mathematisch gesehen gerichtete Graphen mit gewichteten Kanten und können dementsprechend auch als solche dargestellt werden.[466] Ein Graph ist ein mathematisches Modell einer Netzstruktur.[467] Seine Definition ruft daher auch die bereits bekannte Definition eines künstlichen neuronalen Netzes in Erinnerung:

> „Ein (gerichteter) Graph ist ein Paar G = (V, E) bestehend aus einer (endlichen) Menge V von Knoten (engl. vertices, nodes) und einer (endli-

[463] Siehe als Beispiel die Erklärung der Netzarchitektur oben, Kapitel 1 E.
[464] Siehe die Beispiele oben, Kapitel 1 D.
[465] Siehe etwa EP1955228 - Method for Determining Cardiac Output; EP0299876 - Pattern recognition system.
[466] Siehe dazu ausführlich *Kruse u. a.*, Computational Intelligence, S. 33 ff.
[467] *Tittmann*, Graphentheorie: eine anwendungsorientierte Einführung, S. 11.

> chen) Menge $E \subseteq V \times V$ von Kanten (engl. edges, arcs). Wir sagen, dass eine Kante $e = (u, v) \in E$ vom Knoten u auf den Knoten v gerichtet sei.“[468]

Die Knoten des Graphs KNN sind die künstlichen Neuronen, die Kanten die Verbindungen zwischen diesen.
Die Beschreibung als Graph ist ein abstraktes Modell der Struktur des Netzes. Sie enthält daher keine Information darüber, wie die einzelnen Knoten und Kanten beschaffen sind.[469] Das bedeutet, dass die Beschreibung als Graph nichts darüber aussagt, aus welchen mathematischen Funktionen ein künstliches Neuron besteht. Dabei handelt es sich jedoch nur um einen kleinen Teil der Information des KNN. Denn das Repertoire an Formeln und Schwellenwerten der Neuronen ist relativ klein und lässt sich verhältnismäßig leicht aus der zu lösenden Aufgabe schließen. Der weitaus bedeutendere Teil der Information des Netzes, nämlich die Architektur und die Gewichte, können in der Beschreibung als Graph abgebildet werden.
Graphen werden häufig schematisch dargestellt, wobei die Knoten (Neuronen) üblicherweise als Kreise, die Kanten (Verbindungen) als Pfeile dargestellt werden.[470] Die Gewichte der Verbindungen können als Zahlen neben den Pfeilen dargestellt werden (s. **Abb. 5**).

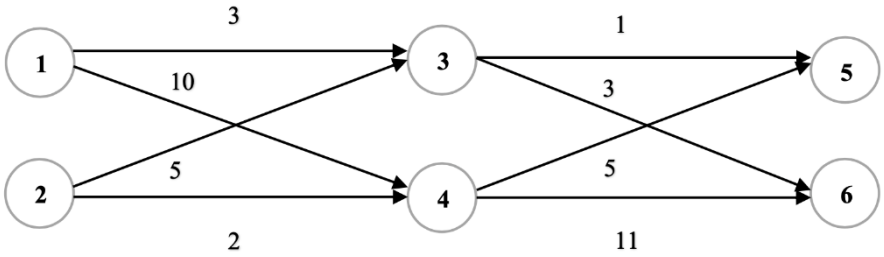

Abb. 5 *Schema eines gerichteten, gewichteten Graphen mit sechs Knoten und acht Kanten*

Die schematische Darstellung eines Graphen inklusive der Gewichte kann für größere Netze schnell unübersichtlich werden und ist daher nur für verhältnismäßig kleine KNN sinnvoll. Hinzu kommt, dass die Information des KNN in

[468] *Kruse u. a.*, Computational Intelligence, S. 33.
[469] *Tittmann*, Graphentheorie: eine anwendungsorientierte Einführung, S. 11.
[470] Diese Darstellung findet sich auch in Patentanmeldungen, siehe etwa EP1955228 - Method For Determining Cardiac Output, Zeichnung 3.

dem Schema eines Graphen weiterhin nicht explizit wird. Selbst in einer verhältnismäßig kleinen Skalierung erschließt sich dem Betrachter anhand der als Zahlenwerte dargestellten gewichteten Verbindungen nicht die Logik einer Entscheidung des Netzes. Es besteht mithin ein erheblicher Unterschied etwa zu Entscheidungsbäumen, in denen die einzelnen Punkte faktische semantische Information repräsentieren.

Selbst im Bereich tiefer neuronaler Netze werden jedoch Graphen als Darstellungsform genutzt. Dann geht es allerdings weniger um eine Darstellung jeder einzelnen gewichteten Verbindung als Zahlenwert, sondern um eine Darstellung der Stärke der Gewichtungen beziehungsweise des Informationsstroms im Netz. Da sie eine weniger exakte, jedoch zugleich anschaulichere Darstellung der Information des Netzes bieten, werden diese Graphen, die in der Forschung häufig als *Node-Link-Diagram* bezeichnet werden, auf der vierten Darstellungsstufe im Bereich XAI erörtert.[471]

Die wohl üblichste Darstellungsform eines KNN als Graph ist die Adjazenzmatrix. Aus dieser lässt sich ablesen, welche Knoten des Netzes durch eine Kante verbunden sind und wie diese Verbindung gewichtet ist.[472] Für den obigen Graphen ergäbe sich folgende Matrix:

	1	2	3	4	5	6
1	0	0	3	10	0	0
2	0	0	5	2	0	0
3	0	0	0	0	1	3
4	0	0	0	0	5	11
5	0	0	0	0	0	0
6	0	0	0	0	0	0

Abb. 6 *Adjazenzmatrix des in Abb. 5 dargestellten Graphen*

[471] Siehe Kapitel 8.

[472] Auch die Netzarchitektur lässt sich mithin in einer Matrix darstellen. So bereits *Springorum*, in: Brunnstein/Sint, Intellectual Property Rights and New Technologies. Proceedings of the KnowRight'95 Conference, S. 209.

Mithilfe einer solchen Matrix ließen sich die gewichteten Verbindungen eines KNN auch in einer Patentanmeldung darstellen, um dem Offenbarungserfordernis Genüge zu tun. Die Darstellungsform scheint jedoch noch nicht verbreitet zu sein, obwohl durch sie eine vom EPA im Einzelfall geforderte Offenlegung der Trainingsdaten wohl vermieden werden könnte.[473] Je nachdem, welche Anforderungen an die Offenbarung und Nacharbeitbarkeit von KNN in Patentanmeldungen in Zukunft gestellt werden, könnte die Darstellung der Gewichte in Adjazenzmatrizen relevant werden.
Für die Repräsentation von Information in einer Adjazenzmatrix gilt das für die schematische Darstellung Gesagte erst recht. Aus ihr lässt sich die Logik des Netzes ebenso wenig herauslesen und selbst der Aufbau des Netzes ist wesentlich weniger intuitiv zu erfassen. Die Information des KNN ist mithin in einer Matrix weiterhin nur implizit repräsentiert. Den Zahlenreihen lässt sich für einen Menschen nicht das „Wissen" des Netzes entnehmen:

> „Even if we read these strings of numbers, we could not relate them to the meaning of knowledge coded in such a way. In other words, we are not able to interpret them in terms of the problem description."[474]

III. Mathematisches Modell

Ein trainiertes künstliches neuronales Netz ist nichts Weiteres als ein mathematisches Modell, also eine in ein Computerprogramm implementierte Berechnungsvorschrift.[475]
Dementsprechend kann es theoretisch auch mathematisch als Formel dargestellt werden.[476] Einen kleinen Ausschnitt einer solchen Darstellung findet sich oben bei der abstrakten Beschreibung eines einzelnen Neurons.[477] Welchen Informationsgehalt die mathematische Darstellung als Formel hat, muss differenziert be-

---

[473] Siehe zum Erfordernis der Offenlegung von Trainingsdaten EPA, 12.5.2020, T 0161/18 – *Äquivalenter Aortendruck/ARC SEIBERSDORF*.

[474] *Flasiński*, Introduction to Artificial Intelligence, S. 225.

[475] *Gesellschaft für Informatik*, Technische und rechtliche Betrachtungen algorithmischer Entscheidungsverfahren, S. 54 f.

[476] Siehe dazu auch *Hauck/Cevc*, ZGE 2019, 135 (144). Für größere Netze ist die Funktion allerdings viel zu komplex, um sich herleiten zu lassen.

[477] Siehe oben Kapitel 1 D; für die vollständige mathematische Darstellung eines Ausgabeneurons eines mehrschichtigen neuronalen Netzes siehe *Thomas/Edelman/Crook*, Credit scoring and its applications, S. 59.

trachtet werden. Solange die Gewichte als Variablen (z. B. w_{ji}) dargestellt werden, entspricht der Informationsgehalt der Formel in etwa derjenigen des Quellcodes. Aus ihr können die Topologie des Netzes und die verwendeten Funktionen abgelesen werden. Natürlich fehlen Informationen, die der Programmierung eigen sind, wie etwa Steuerungsbefehle. Bei diesen handelt es sich jedoch um keine für KNN spezifische Information.
Sind die Werte der Gewichte[478] bereits in die Formel eingesetzt, so ist ihr Informationsgehalt mit dem des Maschinencodes vergleichbar: sie enthält die gesamte semantische Information des Netzes. Mithilfe der Formel ließe sich ein funktionsgleiches Netz in Quellcode entwerfen und in Maschinencode kompilieren.
Die Beziehungen zwischen den Merkmalen und der Einfluss einer Eingabe auf die Ausgabe lassen sich mithin mathematisch darstellen, entziehen sich jedoch je nach Komplexität gänzlich dem menschlichen Verständnis.[479] Die Information des Netzes ist daher in der mathematischen Formel weiterhin nur implizit repräsentiert.

B. Transparenzpflichten und Geheimnisschutz

Die dritte Darstellungsstufe betrifft Informationen, die sich vermeintlich gut für die Erfüllung von Informationspflichten eignen. Die Darstellung etwa eines Graphen scheint auf den ersten Blick, ähnlich einem Entscheidungsbaum, ein gutes Mittel für die Erklärung eines KNN gegenüber einem Laien zu sein.

I. Datenschutzgrundverordnung

Die drei Darstellungsformen könnten mithin grundsätzlich in Frage kommen, um den Informations- und Auskunftsrechten nach Artikel 13, 14 und 15 DSGVO nachzukommen.
Denn zu den Informationen über die involvierte Logik werden teilweise, wie oben dargelegt, auch „Algorithmen“ (darunter könnte wohl untechnisch die mathematische Formel subsumiert werden) sowie Parameter und deren Gewichtungen gezählt.

---

[478] Man spricht auch von Parametrierungen, vgl. *Hauck/Cevc*, ZGE 2019, 135 (144).

[479] Siehe dazu auch *Gesellschaft für Informatik*, Technische und rechtliche Betrachtungen algorithmischer Entscheidungsverfahren, S. 55.

Allerdings ergeben sich auch hier aus den zwei zu beachtenden Parametern Geheimnisschutz und Verständlichkeit erhebliche Einschränkungen.
Denn im Gegensatz zu anderen ML-Verfahren ist die Informationsverarbeitung in den untersuchten Darstellungsformen weiterhin nicht explizit und Verständlichkeit mithin nicht gegeben. Auch lassen sich aus ihnen die Parameter nicht ablesen, was die Darstellung als Graph wesentlich von einem Entscheidungsbaum unterscheidet. Daher ist auch die BGH-Rechtsprechung zu den SCHUFA Scores nur bedingt übertragbar: anders als bei KNN handelt es sich bei der dort betroffenen logistischen Regression um ein vergleichsweise transparentes und mithin nicht um ein Blackbox-Modell.[480] Die Informationsverarbeitung ist explizit repräsentiert und kann für jedermann nachvollziehbar dargestellt werden. Die Offenlegung der Parameter und ihrer Gewichtungen sowie der Berechnungen eines Regressions-Modells ist mithin verhältnismäßig einfach und führt auch automatisch zur Offenlegung des Geschäftsgeheimnisses. Die mathematische Berechnungsvorschrift hingegen, auf der das KNN beruht, kann so komplex sein kann, dass sie von einem menschlichen Betrachter nicht mehr nachvollzogen werden kann.[481]
Bei KNN gehen mithin Offenbarung des Geschäftsgeheimnisses und Transparenz des Entscheidungsmechanismus nicht Hand in Hand. Im Gegenteil: durch Offenlegung sowohl der Beschreibung des Netzes in natürlicher Sprache und seiner Adjazenzmatrix, oder auch durch Offenlegung der Berechnungsvorschrift mit integrierten Gewichten, würde das Geschäftsgeheimnis offenbart. Aufgrund der Opazität der Information könnte der Informationspflicht gegenüber einem Verbraucher auf diese Weise jedoch nicht nachgekommen werden.
Dies gilt umso mehr, als Transparenzpflichten regelmäßig KNN zum Gegenstand haben werden, die Netze wie das in den Abb. 5 und 6 dargestellte KNN um ein Vielfaches überschreiten, und deren Lösungsweg mithin nicht nachvollziehbar ist. Daher geht auch die Gesellschaft für Informatik in ihrem Gutachten im Auftrag des Sachverständigenrats für Verbraucherfragen davon aus, dass

> „hochdimensionale Matrizen[,] die Gewichte und Assoziationen zwischen den Neuronen und Schichten darstellen, [...] zur Interpretation durch einen Menschen bei Netzen nichttrivialer Komplexität im Regel-

[480] *Gesellschaft für Informatik*, Technische und rechtliche Betrachtungen algorithmischer Entscheidungsverfahren, S. 53.
[481] *Gesellschaft für Informatik*, Technische und rechtliche Betrachtungen algorithmischer Entscheidungsverfahren, S. 55.

> fall nicht geeignet [sind] und [...] als Erklärungskomponente dadurch nicht in Frage [kommen]." [482]

Die Darstellungsformen der dritten Stufe sind daher für die Erfüllung der Transparenzverpflichtungen aus der DSGVO kaum geeignet: einerseits muss der Parameter der Verständlichkeit bei ihnen verneint werden, andererseits führt ihre Offenlegung zu einer Offenbarung des Geschäftsgeheimnisses am KNN. Die erforderliche Interessenabwägung muss daher, unabhängig von ihrer dogmatischen Anknüpfung, zugunsten des Geheimnisinhabers ausfallen. Denn eine Offenlegung von Information, welche das Verständlichkeitsgebot des Artikel 12 Abs. 1 DSGVO nicht erfüllt und zugleich die Offenbarung von Geschäftsgeheimnissen zur Folge hat, ist nicht zu rechtfertigen.
Im Rahmen der Auskunfts- und Informationsrechte der DSGVO spielen daher die im folgenden Kapitel auf der vierten Darstellungsstufe aufgezeigten Möglichkeiten der Darstellung der Information eines KNN eine herausragende Rolle. Sie könnten die einzige Möglichkeit darstellen, denjenigen, die von ADM durch ein KNN betroffen sind, die notwendige Information zur Verfügung zu stellen, um sich gegen eine möglicherweise rechtswidrige Datenverarbeitung zu wehren.

II. Entwurf der KI-Verordnung

Anders verhält es sich bei den Transparenzpflichten, die der KI-VO-E vorsieht. Da diese grundsätzlich[483] nicht gegenüber Verbrauchern, sondern gegenüber Nutzern und Experten bestehen, spielt der Parameter der Verständlichkeit dort, wenn überhaupt, nur eine untergeordnete Rolle.

1. Private Informationsempfänger

Ob den Nutzern mit der Gebrauchsanweisung des KI-Systems (Art. 13 Abs. 2 KI-VO-E) auch Information über die Netzarchitektur und die Gewichte zur Verfügung gestellt werden muss, kann noch nicht abschließend beantwortet

---

[482] *Gesellschaft für Informatik*, Technische und rechtliche Betrachtungen algorithmischer Entscheidungsverfahren, S. 44.

[483] Die Transparenzpflicht aus Art. 52 KI-VO-E spielt hier keine Rolle, da mit ihr keine Offenlegung von geheimer Information verbunden sein dürfte. Siehe dazu und zum Adressatenkreis der Gebrauchsanweisung oben, Kapitel 3 C 2.

werden. Zumindest für die Gewichte dürfte eine so weitgehende Offenlegung jedoch äußerst fraglich sein. Immerhin muss die Information in der Gebrauchsanweisung nicht nur „präzise" und „vollständig", sondern eben auch „in einer für die Nutzer relevanten [...] und verständlichen Form" zur Verfügung gestellt werden gem. Art. 13 Abs. 2 KI-VO-E. Zudem soll das System lediglich „angemessen" verstanden und verwendet werden können, was den Umfang der erforderlichen Transparenz ebenfalls einschränkt.
Zur Erfüllung einer derart weit ausgelegten Pflicht wären jedenfalls alle in diesem Kapitel aufgezeigten Darstellungsformen geeignet, aus denen sich beides ergibt, also insbesondere die Darstellung als Matrix und Formel oder – bei Netzen kleiner Größe – auch als Graph. Die Beschreibung in natürlicher Sprache eignet sich weniger für die Darstellung der Gewichte, erscheint jedoch gerade für die Gebrauchsanweisung als gegebenenfalls ergänzende Information gut geeignet.
Im Hinblick auf die Nutzer als Empfängergruppe ist die Bereitstellung derart weitgehender Information über ein KNN jedoch problematisch. Denn die Information der dritten Darstellungsstufe lässt eine Reproduktion eines funktionsgleichen Netzes zu. Dies gilt umso mehr, als der Nutzer im Sinne des KI-VO-E naturgemäß genau über die Einsatzzwecke des Netzes und auch zwingend über die erforderliche Aufbereitung der Eingabedaten Bescheid weiß.[484]
Wie oben bereits dargelegt, verpflichtet der KI-VO-E auch die Nutzer zur Wahrung der Vertraulichkeit, was jedoch die Gefahr des faktischen Geheimnisverlusts fortbestehen lässt. Unklarheiten im Umfang der Offenlegungspflichte und hinsichtlich des Verhältnisses von Transparenz und Geheimnisschutz sind vor diesem Hintergrund äußerst problematisch.
Im folgenden Kapitel wird daher untersucht, ob die Darstellungsformen der vierten Stufe eine Möglichkeit bieten, den Transparenzpflichten mit einem ähnlichen Informationsgehalt für den Nutzer und zugleich mit geringerem Risiko für die Anbieter als Inhaber von Geschäftsgeheimnissen nachzukommen.

2. Öffentliche Informationsempfänger

Die im KI-VO-E vorgesehenen Offenlegungspflichten gegenüber zuständigen nationalen Behörden und anderen öffentlichen Stellen sowie gegenüber notifizierten Stellen sind sowohl hinsichtlich ihres Umfangs als auch hinsichtlich des Schutzes von Geschäftsgeheimnissen klarer ausgestaltet.

[484] Siehe nur den gem. Art. 13 Abs. 3 lit. b) v) und vi) geforderten Inhalt der Gebrauchsanweisung.

Notifizierten Stellen ist, wie gezeigt, im Falle einer externen Konformitätsbewertung die technische Dokumentation zur Prüfung vorzulegen (Anhang VII Ziff. 4 des KI-VO-E). Diese dürfte sowohl Information über die Architektur als auch über die Gewichte eines trainierten KNN enthalten müssen (Anhang IV Ziff. 2b). Die in diesem Kapitel aufgeführten Darstellungsmöglichkeiten von KNN sind geeignet, um diese Informationen in der technischen Dokumentation darzustellen. Die Beschreibung in natürlicher Sprache und die schematische Darstellung als Graph eignen sich zur Darstellung der „allgemeinen Logik" kleinerer Netze, während etwa Matrizen die Gewichte größerer Netze entnommen werden können. Der Pflicht zur Dokumentation des „Algorithmus" (Anhang IV Ziff. 2b) kann durch die Darstellung des KNN als mathematische Formel nachgekommen werden.

In allen Darstellungsformen der dritten Stufe werden Geschäftsgeheimnisse offenbar. Dies ist jedoch gegenüber den notifizierten Stellen, die wie gezeigt Verschwiegenheitspflichten unterliegen sollen, weniger riskant als gegenüber der großen Gruppe der Nutzer. Dennoch verbleibt aufgrund des rein faktischen Schutzes des GeschGehG immer ein Restrisiko des Geheimnisverlustes, das mit jeder Erweiterung des Empfängerkreises weiter wächst. Die den notifizierten Stellen eingeräumte Möglichkeit, ihrerseits bestimmte Aufgaben an Zweigstellen zu übertragen oder an Unterauftragsnehmer zu vergeben (Art. 34 KI-VO-E), kann in diesem Zusammenhang als kritisch bewertet werden.

Ob angesichts diese Restrisikos eine andere Möglichkeit besteht, den sich aus dem KI-VO-E gegenüber notifizierten Stellen bestehenden Informationspflichten nachzukommen, wird auf der vierten Darstellungsstufe im nachfolgenden Kapitel untersucht.

Die im KI-VO-E vorgesehenen Offenlegungspflichten gegenüber zuständigen nationalen Behörden sind naturgemäß die Umfangreichsten. Auch ihnen kann durch die Darstellungsmöglichkeiten der dritten Stufe nachgekommen werden. Für die Überprüfung der Funktionsweise des KNN durch Marküberwachungsbehörden wird jedoch der Zugang zum Quellcode in Kombination mit dem Zugang zu Trainings- und Testdatensätzen (vgl. Art. 63 Abs. 8 und 9 KI-VO-E), oder auch der Zugang zum Maschinencode, wesentlich sinnvoller sein als der Zugang zu den Informationen der dritten Darstellungsstufe, anhand derer sie das Netz zunächst nachbauen müssten, um es anschließend überprüfen zu können.

Nichtsdestotrotz dürfte eine mögliche Offenbarung von Graphen, Beschreibungen oder der mathematischen Formel, etwa durch Übergabe der techni-

schen Dokumentation gemäß Art. 11 KI-VO-E, gegenüber Behörden und öffentlichen Stellen für die Geheimnisinhaber aufgrund der im KI-VO-E gesondert normierten und für Beschäftigte des öffentlichen Dienstes ohnehin bestehenden Verschwiegenheitspflichten das geringste Risiko bergen.[485] Dennoch stellt auch hier die vorgesehene Ausweitung des Empfängerkreises auf den Grundrechtsschutz überwachende oder durchsetzende öffentliche Stellen und Behörden ein erhöhtes Risiko für den Anbieter und Geheimnisinhaber dar.
Auch für den Empfängerkreis der zuständigen nationalen Behörden gemäß Art. 2 Nr. 43 KI-VO-E (mit Ausnahme der Marktüberwachungsbehörde) und die öffentlichen Stellen wird daher im nächsten Kapitel untersucht, ob den Offenlegungspflichten aus dem KI-VO-E durch Darstellungsformen der vierten Stufe nachgekommen werden kann.

C. Urheberrechtlicher Schutz

I. Beschreibung mittels natürlicher Sprache

Die Beschreibung von Struktur und verwendeten Formeln eines KNN in natürlicher Sprache enthält grundsätzlich dieselbe semantische Information über das KNN wie der Quellcode.[486] Auch mit ihrer Offenlegung würde daher nur ein Teil des Geschäftsgeheimnisses erlangt, der jedoch die Erlangung des gesamten Geheimnisses begünstigen könnte. Daher steht auch hier die Frage einer weitergehenden Schutzmöglichkeit im Raum.
Ein Schutz als Computerprogramm gemäß § 69a UrhG scheidet mangels Steuerungsbefehlen bei der sprachlichen Umschreibung aus. Zu denken könnte allenfalls an einen Schutz als Entwurfsmaterial sein, das gemäß § 69a Abs. 1 UrhG auch in den Schutzbereich einbezogen ist. Als Entwurfsmaterial gelten alle „Produkte des Entwicklungsprozesses" eines Computerprogramms.[487] Voraussetzung ist gemäß Erwägungsgrund 7 der Computerprogramm-Richtlinie[488], dass

[485] Für Beamte ist die in Artikel 33 Abs. 5 GG verankerte Verschwiegenheitspflicht einfachgesetzlich in § 67 Bundesbeamtengesetz und § 37 Beamtenstatusgesetz normiert.

[486] Abgesehen von im Quellcode enthaltenen Zusatzinformationen, die für den semantischen Gehalt des KNN keine Relevanz haben.

[487] *Wiebe*, in: Spindler/Schuster/Anton, Recht der elektronischen Medien, § 69a Rn. 7.

[488] Richtlinie 2009/24/EG des Europäischen Parlaments und des Rates vom 23. April 2009 über den Rechtsschutz von Computerprogrammen.

die durch sie geleistete „Art der vorbereitenden Arbeit die spätere Entstehung eines Computerprogramms zulässt."
Für die sprachliche Beschreibung eines KNN wird man jedoch von einer solchen vorbereitenden Arbeit für die spätere Entstehung des Computerprogramms nicht ausgehen können. Mit ihr wird nicht das Computerprogramm selbst, sondern die ihm zugrundeliegende Idee umschrieben, die dann in einem Computerprogramm implementiert wird und die gemäß § 69a Abs. 2 S. 2 UrhG nicht geschützt ist.[489]
Es könnte allenfalls ein Schutz als Sprachwerk gemäß § 2 Abs. 1 Nr. 1 UrhG in Betracht kommen. Auch hier gilt jedoch: der Schutzumfang wäre auf die konkrete Gestaltung der Beschreibung beschränkt, die enthaltenen Informationen über die Netzarchitektur und die verwendeten Formeln blieben nach den bereits zu § 69a Abs. 1 S. 2 UrhG dargestellten Grundsätzen frei.[490]

II. Darstellung als Graph

Zur Offenlegung der Information des als schematischer Graph, als Kantenliste oder als Adjazenzmatrix dargestellten KNN gelten die Ausführungen zu den Gewichten als Datei entsprechend. Denn wird die Datei in einem für Menschen lesbaren Format gespeichert, so erfolgt die Darstellung meist als Matrix. Der Informationsgehalt der drei Darstellungsformen ist derselbe, auch wenn die schematische Darstellung sich ab einer gewissen Größe weniger eignet. Auch hier stellt sich mithin die Frage des urheberrechtlichen Schutzes.
Bei der Darstellung als Graph könnte man ebenfalls an einen Schutz als Entwurfsmaterial gemäß § 69a Abs. 1 UrhG denken. Die erforderliche Nähe zum angestrebten Code wird jedoch wie bei der sprachlichen Beschreibung zu verneinen sein. Allenfalls sofern die Darstellung des KNN gleich einem Entscheidungsbaum bereits qualitativ an einen Programmablaufplan heranreichte, könnte ein Schutz als Entwurfsmaterial in Betracht kommen.[491] Dies wird jedoch meist nicht der Fall sein und der Entscheidungsbaum daher nur die Idee des KNN verkörpern, die als solche grundsätzlich frei bleiben soll.

---

[489] Die Schutzfähigkeit der „Idee der Topologie" als Entwurfsmaterial verneinen auch *Hartmann/Prinz*, in: Taeger, Rechtsfragen digitaler Transformationen - Gestaltung digitaler Veränderungsprozesse durch Recht, S. 779 f.

[490] So auch *Springorum*, in: Brunnstein/Sint, Intellectual Property Rights and New Technologies. Proceedings of the KnowRight'95 Conference, S. 209.

[491] *Söbbing*, MMR 2021, 111 (114).

Die Schutzfähigkeit der Gewichte in Form einer Kantenliste oder Adjazenzmatrix beurteilt sich ebenso wie ihr Schutzfähigkeit als Datei: sie können mangels Unabhängigkeit nicht als Datenbank im Sinne des § 87a UrhG geschützt werden.
Sowohl für die schematische Darstellung als auch für Kantenliste und Matrix käme gegebenenfalls auch ein Schutz als Darstellung wissenschaftlicher oder technischer Art gemäß § 2 Abs. 1 Nr. 7 UrhG in Betracht. Auch nach dieser würde jedoch allenfalls die konkrete Ausgestaltung geschützt sein, die semantische Information bliebe frei.[492]

III. Mathematisches Modell

Die mathematische Darstellung eines KNN als Formel trägt, in Abhängigkeit von der Integration der Gewichte, dieselbe Information über das KNN wie der Quellcode bzw. der Maschinencode. Mit ihrer Offenlegung würde mithin auch das Geschäftsgeheimnis am KNN ganz oder teilweise erlangt.
Auch für das mathematische Modell käme eine Schutzfähigkeit allenfalls als Entwurfsmaterial gemäß § 69a Abs. 1 UrhG in Betracht. Dies wird jedoch vor dem Hintergrund zu verneinen sein, dass die Erzeugung des Quellcodes aus dem mathematischen Modell nicht ohne größere Zwischenschritte möglich ist und mithin auch hier die erforderliche Nähe zum zu entwickelnden Code nicht gegeben ist.[493]

IV. Fazit

Der urheberrechtliche Schutz der Darstellungsformen der dritten Stufe ist mithin, sofern er überhaupt angenommen werden kann, lückenhaft. Der reine Schutz der syntaktischen Information bewahrt den Geheimnisinhaber nicht vor Erlangung und widerrechtlicher Nutzung seines Geschäftsgeheimnisses, der semantischen Information am KNN.
Im folgenden Kapitel werden daher weitere Darstellungsmöglichkeiten auf ihr Potenzial hin untersucht, die Entscheidungsfindung eines KNN zu erklären, ohne gleichzeitig dessen semantische Information zu offenbaren.

---

[492] So auch *Springorum*, in: Brunnstein/Sint, Intellectual Property Rights and New Technologies. Proceedings of the KnowRight'95 Conference, S. 209.
[493] *Söbbing*, CR 2020, 223 (228).

Kapitel 8

Vierte Darstellungsstufe: Explainable Artificial Intelligence

In diesem Kapitel wird der Frage nachgegangen, ob Methoden der Explainable Artificial Intelligence die Dichotomie zwischen Transparenz und Geheimnisschutz auflösen können. Möglicherweise kann mit ihrer Hilfe den analysierten Transparenzpflichten unter Wahrung von Geschäftsgeheimnissen nachgekommen werden.

A. Einführung Explainable Artificial Intelligence

An der Schnittstelle von Technik und Recht ist die Frage der Erklärbarkeit Künstlicher Intelligenz von großer Relevanz.
Während die GOFAI als symbolische KI durch ihre explizite Wissensrepräsentation aus sich heraus erklärbar ist,[494] stellt die implizite Repräsentation von Information in KNN viele Anwendungsfälle in ganz unterschiedlichen Disziplinen vor ein Problem. Denn die Praxis ist häufig auf eine symbolhafte Darstellung von ML-Modellen angewiesen. Am intensivsten beschäftigt sich bislang wohl das Datenschutzrecht und allgemein das Verbraucherschutzrecht mit der Frage der Erklärbarkeit.[495] Aber auch für andere Disziplinen ist sie sehr bedeutend: Unterstützt Maschinenlernen beispielsweise die Entscheidungsfindung eines Strafrichters, ist die Transparenz des eingesetzten Systems für Angeklagte, Verteidiger, Richter einer nächsten Instanz, aber auch für Opfer enorm wichtig. Auch in der Medizin und im Haftungsrecht stellt sich die Frage nach der Nach-

[494] Siehe dazu auch schon *Andrews/Diederich/Tickle*, Knowledge-Based Systems 1995, 373 (374).

[495] So zum Datenschutzrecht *Hacker/Krestel/Grundmann/Naumann*, Artificial Intelligence and Law 2020, 415 (416); siehe zu Techniken der sog. „rule extraction" für Kredit-Scoring-Modelle *Thomas/Edelman/Crook*, Credit scoring and its applications, S. 70 ff.

vollziehbarkeit von Entscheidungen durch Systeme künstlicher Intelligenz mit Nachdruck.[496] Von besonderer Bedeutung ist die Erklärbarkeit auch für den Einsatz künstlicher Intelligenz im militärischen Bereich.[497] Intransparenz und mangelnde Erklärbarkeit stellen mithin eine der großen Herausforderungen für die breite Anwendung von ML-Anwendungen dar.[498]

Transparenz- und Regulierungsbestrebungen sind jedoch keineswegs der hauptsächliche Antrieb für die Suche nach Erklärbarkeit. Auch Entwickler und Forscher haben ein intrinsisches Interesse daran, die Funktionsweise ihrer Schöpfungen nachzuvollziehen. Denn diese Nachvollziehbarkeit ermöglicht es, Modelle zu verbessern und besonders gute Hyperparameter zu finden. Außerdem macht sie den Einsatz von KNN für ein breiteres Spektrum von Anwendern interessanter: wer ein fertiges KNN für seine Zwecke einsetzen möchte, muss nicht mehr über tiefgehende Informatikkenntnisse verfügen, um die grundlegende Funktionsweise seines Modells zu verstehen.[499]

Transparenz kann daher unterschiedliche Ziele und Zielgruppen haben: sie kann Entwicklern bei Problemlösungen helfen, Vertrauen des Anwenders und der Gesellschaft in ML-Techniken an sich stärken oder die Akzeptanz einer konkreten Entscheidung fördern sowie Audits und Testungen und damit auch Regulierung erlauben.[500]

Bereits 1996 sahen Andrews et al. die Fähigkeit zur Erklärung als „Vehikel, um die Grenze zwischen konnektionnistischen und symbolischen Ansätzen zu verbinden."[501] Die Suche nach Erklärbarkeit von ML-Modellen hat dann zur Herausbildung eines eigenständigen Forschungszweigs geführt, der sog. *Explainable Artificial Intelligence*.[502] Für die Anwender von ML-Modellen über

---

[496] *Hacker/Krestel/Grundmann/Naumann*, Artificial Intelligence and Law 2020, 415 (419 ff.).

[497] *Knight*, MIT Technology Review 2017, 53 (61).

[498] *Samek/Müller*, in: Samek u. a., Explainable AI: Interpreting, Explaining and Visualizing Deep Learning, S. 6.

[499] *Yosinski u. a.*, Deep Learning Workshop, 31st International Conference on Machine Learning 2015, 1 (2 f.).

[500] Diese und weitere Ziele nennt *Weller*, in: Samek u. a., Explainable AI: Interpreting, Explaining and Visualizing Deep Learning, S. 25; ähnlich auch schon *Andrews/Diederich/Tickle*, Knowledge-Based Systems 1995, 373 (374 f.).

[501] „It [the addition of an explanation capability] also provides a vehicle for traversing the boundary between the connectionist and symbolic approaches." *Andrews/Diederich/Tickle*, Knowledge-Based Systems 1995, 373 (373 f.).

[502] Teilweise wird auch von „AI neuroscience" gesprochen, vgl. *Voosen*, How AI detectives are cracking open the black box of deep learning, Science, http://www.science-

die unterschiedlichen Disziplinen hinweg kann XAI im besten Fall helfen, Fehler in einem Modell aufzudecken und zu korrigieren.[503]
Die Disziplin der Expainable Artificial Intelligence verfügt über keine einheitliche Terminologie.[504] Vor allem die wesentlichen Begriffe Erklärbarkeit (*explainability*), Interpretierbarkeit (*interpretability*) und Transparenz (*transparency*) werden unscharf und uneinheitlich gebraucht.[505]
Auch besteht kein Konsens darüber, was überhaupt eine Erklärung eines ML-Modells ausmacht.[506] Eine gute Beschreibung des gewünschten Inhalts einer Erklärung eines KNN geben Andrews et al.:

> "[...] within a trained artificial neural network, knowledge acquired during the training phase is encoded as (a) the network architecture itself (e.g. the number of hidden units), (b) an activation function associated with each (hidden and output) unit of the ANN, and (c) a set of (real-valued) numerical parameters (called weights). In essence, the task of extracting explanations (or rules) from a trained artificial neural network is therefore one of interpreting in a comprehensible form the collective effect of a, b and c."[507]

Danach enthält eine Erklärung eine verständliche Interpretation der gemeinsamen Wirkung der unterschiedlichen Bestandteile des ML-Modells. Diese Definition trifft noch keine Entscheidung darüber, auf welcher Ebene des Modells die Erklärung ansetzt. Unabhängig von der gewählten Terminologie werden nämlich üblicherweise Modelle unterschieden, die für den Menschen von sich heraus verständlich sind und Methoden, die eine konkrete Entscheidung eines

---

mag.org/news/2017/07/how-ai-detectives-are-cracking-open-black-box-deep-learning (zuletzt abgerufen am 26.10.2023).

[503] *Samek/Müller*, in: Samek u. a., Explainable AI: Interpreting, Explaining and Visualizing Deep Learning, S. 8.

[504] „A theory of explainable AI, with a formal and universally agreed definition of what explanations are, is lacking.", *Samek/Müller*, in: Samek u. a., Explainable AI: Interpreting, Explaining and Visualizing Deep Learning, S. 17.

[505] Siehe nur die Differenzen zwischen *Russell/Norvig*, Artificial intelligence, S. 711 f.; *Lipton*, ICML WHI 2016, 96 (98 ff.).

[506] So auch *Guidotti u. a.*, ACM Computing Surveys 2019, 1 (36).

[507] *Andrews/Diederich/Tickle*, Knowledge-Based Systems 1995, 373 (375 f.).

für den Menschen unverständlichen Modells erklären können.[508] Die letztgenannten Methoden bilden den Schwerpunkt der Forschung zu XAI in den letzten Jahren.[509]

In unterschiedlichen Domänen gibt es zwar auch Bestrebungen, von Anfang an interpretierbare sog. Whitebox-Modelle zu verwenden, statt Blackbox-Modelle im Nachhinein mithilfe von XAI-Methoden zu durchleuchten.[510] Da jedoch der Zielkonflikt zwischen Exaktheit und Interpretierbarkeit (*tradeoff between accuray and interpretability*) fortdauert, kann nicht auf eine schnelle Ablösung von Blackbox-Modellen durch Whitebox-Modelle gesetzt werden.[511] Daher nimmt die vorliegende Arbeit XAI-Ansätze in den Fokus, die Blackbox-Modelle erklärbar machen sollen.

Hier wird im Folgenden die von Lipton verwendete Terminologie und Taxonomie gebraucht, die besonders deutlich erscheint.[512] Lipton unterscheidet einerseits Modelle, deren Entscheidungsmechanismus aus einer ex-ante-Sicht nachvollzogen werden kann und nennt diese „transparent". Lediglich „post-hoc-in-

[508] Diese Einteilung ist weit verbreitet. Russell/Norvig sprechen von „interpretable" und „explainable" Methoden: *Russell/Norvig*, Artificial intelligence, S. 711 f.; Holzinger von „Ante-hoc-" und „Post-hoc-"Ansätzen: *Holzinger*, Informatik-Spektrum 2018, 138 (140 ff.); Lipton spricht von „transparency" und „post-hoc interpretability": *Lipton*, ICML WHI 2016, 96 (98 ff.); Doshi-Velez/Kim sprechen von „global" und „local interpretability": *Doshi-Velez/Kim*, arXiv:1702.08608, 2017, 1 (7); ebenso *Guidotti u. a.*, ACM Computing Surveys 2019, 1.

[509] So *Guidotti u. a.*, ACM Computing Surveys 2019, 1 (26); *Adadi/Berrada*, IEEE Access 2018, 52138 (52147).

[510] Siehe allgemein zur Forderung *Rudin*, Nature Machine Intelligence 2019, 206; siehe für Beispiele von Whitebox-Modellen *Adadi/Berrada*, IEEE Access 2018, 52138 (52147); siehe im Bereich Kreditscoring die Patentanmeldung „Optimizing Neural Networks for generating analytical or predictive outputs" der Firma Equifax Inc., US 2018/0025273 A1.

[511] Siehe zum „Tradeoff" nur *Adadi/Berrada*, IEEE Access 2018, 52138 (52142); *Hacker/Krestel/Grundmann/Naumann*, Artificial Intelligence and Law 2020, 415 (416).

[512] Eine andere Systematisierung von XAI-Ansätzen nehmen etwa Gilpin et al. vor und trennen Methoden, welche die Darstellung von Daten oder der Datenverarbeitung im Netz erklären wollen, von solchen, die selbst eine Erklärung ihrer Entscheidung kreieren. Siehe dazu *Gilpin u. a.*, arXiv:180600069 2019, 1; für eine differenzierte Einteilung nach dem Inhalt der Erklärung siehe *Samek/Müller*, in: Samek u. a., Explainable AI: Interpreting, Explaining and Visualizing Deep Learning, S. 10 f.; eine umfassende Übersicht der unterschiedlichen XAI-Methoden findet sich bei *Adadi/Berrada*, IEEE Access 2018, 52138.

terpretierbar“ nennt er andererseits Modelle, bei denen allenfalls eine konkrete Entscheidung im Nachhinein erklärt werden kann.[513]
Transparenz kann nach Lipton als Gegenteil von Opazität oder „*blackbox-ness*“ verstanden werden und bedeutet daher, dass ein Modell verstanden werden kann.[514] Die so verstandene Transparenz hat viel mit Komplexität des Modells, aber auch mit Skalierung zu tun. Ein dem Grunde nach transparentes Modell, das auf einer logistischen Regression beruht, kann ab einer gewissen Größe nicht mehr nachvollzogen werden. Zudem spielt eine Rolle, ob die durch das Modell gelernten Konzepte an sich dem menschlichen Denken zugänglich sind.[515] Denn die im Rahmen des Trainings eines ML-Modells gelernten Muster werden ohne Ansehung des menschlichen Verständnisses gebildet:

> „When a computer learns and consequently builds its own representation of a classification decision, it does so without regard for human comprehension.”[516]

Lipton unterscheidet weiter zwischen Transparenz des gesamten Modells (*simulatability*), Transparenz einzelner Bestandteile (*decomposability*) und Transparenz des Lernalgorithmus (*algorithmic transparency*).[517] *Simulatability* liegt vor, wenn ein Mensch das Modell an sich erfassen kann. Sie ist gegeben, wenn die Berechnung des Modells innerhalb einer „angemessenen“ Zeitspanne anhand der Eingaben und Parameter nachvollzogen werden kann, was bei (linearen) Modellen mit niedrigem Komplexitätsgrad erfüllt sein dürfte, jedoch bereist nicht mehr bei komplexeren linearen Modellen oder tieferen Entscheidungsbäumen.[518] *Decomposability* ist gegeben, wenn alle Parameter, Eingaben und Berechnungen eines Modells intuitiv verständlich sind, wenn also etwa die Knotenpunkte eines Entscheidungsbaums jeweils nachvollziehbar ein Merkmal

---

[513] *Hacker/Krestel/Grundmann/Naumann*, Artificial Intelligence and Law 2020, 415 (417); *Lipton*, ICML WHI 2016, 96 (98).

[514] *Lipton*, ICML WHI 2016, 96 (98).

[515] *Hacker/Krestel/Grundmann/Naumann*, Artificial Intelligence and Law 2020, 415 (435); Burrell spricht von einem „mismatch between mathematical procedures of machine learning algorithms and human styles of semantic interpretation“, *Burrell*, Big Data & Society 2016, 1 (2).

[516] *Burrell*, Big Data & Society 2016, 1 (10); siehe jedoch für eine Konzept-basierte Erklärung von KNN *Yeh u. a.*, arXiv:1910.07969v5 2020, 1.

[517] Siehe zum Folgenden *Lipton*, ICML WHI 2016, 96 (98).

[518] *Lipton*, ICML WHI 2016, 96 (98).

repräsentieren.[519] *Algorithmic transparency* liegt vor, wenn der Lernprozess nachvollzogen werden kann. Sie dürfte bei einfachen linearen Modellen vorliegen, jedoch nicht bei *deep learning*.[520]
Die hier untersuchten KNN sind an diesen Maßstäben gemessen intransparent. Denn selbst bei kleinen KNN, deren Berechnung anhand der Gewichte noch nachvollzogen werden könnte, ist keine *decomposability* gegeben, da die einzelnen Neuronen in den verborgenen Schichten üblicherweise keine nachvollziehbaren Merkmale abbilden.
Bei den lediglich post-hoc interpretierbaren Modellen unterscheidet Lipton die Methoden sprachliche Erklärung (*text explanations*), Visualisierung (*visualization*) und Erklärung durch Beispiele (*explanation by example*). Diese ex-post-Methoden können die durch ein Modell getroffene Vorhersage erklären, jedoch ohne dass sich daraus die Funktionsweise des Modells ablesen ließe.[521] Es handelt sich also um klassische XAI-Ansätze, an denen nach Liptons Einordnung gut sichtbar wird, dass sich Interpretierbarkeit und „Blackbox"-Charakter nicht zwingend ausschließen. Ein Blackbox-Modell ist zwar nicht transparent im beschriebenen Sinne, die von ihm getroffene Entscheidung lässt sich jedoch durchaus interpretieren.[522]
Text explanations können eine konkrete durch ein Modell getroffene Vorhersage verbal erklären, etwa indem ein zweites neuronales Netz die Vorhersage eines ersten erläutert.[523] Durch *visualization* können diejenigen Parameter herausgestellt werden, auf denen die Entscheidung des Modells hauptsächlich beruht. Bei der *explanation by example* werden weitere Beispiele genannt, die das Modell als dem konkreten Fall ähnlich einstuft.[524]
Zum Typ der post-hoc interpretierbaren Modelle gehören nicht-lineare Methoden mit einer unüberschaubar großen Zahl von Parametern, wie etwa komplexe neuronale Netze, die auf *deep learning* basieren. Doch auch komplexe lineare Modelle können sich einer Transparenz im Sinne Liptons verschließen und teilweise selbst für die Methoden der *post-hoc-interpretability* nicht zugänglich sein.[525]

---

[519] *Lipton*, ICML WHI 2016, 96 (98).
[520] *Lipton*, ICML WHI 2016, 96 (98).
[521] *Lipton*, ICML WHI 2016, 96 (97).
[522] *Lipton*, ICML WHI 2016, 96 (97).
[523] Siehe zu diesem Absatz *Lipton*, ICML WHI 2016, 96 (99).
[524] *Lipton*, ICML WHI 2016, 96 (99).
[525] *Lipton*, ICML WHI 2016, 96 (99).

In diesem Kapitel werden verschiedene XAI-Methoden dargestellt, die sich dem Bereich Visualisierung und Erklärung durch Beispiele zuordnen lassen.
Eine weitere für die vorliegende Forschungsfrage relevante Unterscheidung, die auch bei der Analyse der Transparenzpflichten in Kapitel 3 bereits eine Rolle spielte, ist die nach dem Empfänger einer Erklärung. Diese Unterscheidung wird auch in der XAI-Forschung vorgenommen. So differenzieren etwa Gilpin et al. bei Erklärungsansätzen für tiefe neuronale Netze zwischen an Experten gerichteten technischen Erklärungen (*inside* explanations) auf der einen und gesellschaftsrelevanten Erklärungen des „Warum" einer Entscheidung (*outside* explanations) auf der anderen Seite.[526]

B. XAI-Techniken und Darstellung der Information

Die Forschung zu XAI ist äußerst umfangreich und hat unzählige Methoden zur Erklärung von ML-Modellen hervorgebracht, die darüber hinaus nach den unterschiedlichsten Kriterien klassifiziert werden.[527]
Für die Frage der Offenlegung von Geschäftsgeheimnissen ist jedoch nicht die jeweilige informationstechnische Methode von Interesse, sondern allein ihr Ergebnis. Daher werden hier einige weit verbreitete Darstellungen analysiert, die durch XAI-Techniken hervorgebracht werden. Dabei wird der Fokus zudem auf solche Darstellungen gelegt, die zur Erfüllung etwaiger Transparenzverpflichtungen gegenüber Laien geeignet erscheinen. Denn die Ergebnisse der XAI-Methoden werden sich ganz überwiegend an diese Zielgruppe richten.

I. Heatmaps und Feature Visualization

Visualisierungen sind eine geeignete Methode für die Erklärung hochkomplexer Modelle.[528] Dementsprechend gibt es eine ganze Reihe von Methoden, mit deren Hilfe das Wissen eines KNN bildlich dargestellt werden kann. Drei verbreitete Techniken, die vornehmlich zur Klassifizierung von Bildern eingesetzt werden, sollen hier zusammengefasst vorgestellt werden.

---

[526] *Gilpin/Testart/Fruchter/Adebayo*, arXiv:1901.06560v1 2019, 1 (1 f.).

[527] Einen systematischen Überblick über die aktuelle Forschung geben *Adadi/Berrada*, IEEE Access 2018, 52138.

[528] So auch *Hacker/Krestel/Grundmann/Naumann*, Artificial Intelligence and Law 2020, 415 (433).

Eine Methode zur Visualisierung der Information eines KNN sind sogenannte *Heatmaps* („Wärmekarten").[529] Ein *Heatmap* kann über die zu klassifizierende Eingabe, üblicherweise ein Bild, gelegt werden und diejenigen Regionen farblich hervorheben, die für die Klassifizierung maßgeblich waren. Durch eine farbliche Abstufung in Anlehnung an Bilder einer Wärmekamera kann so auch der Einflussgrad der jeweiligen Region auf die Klassifizierung abgebildet werden.[530]
Ein *Heatmap* kann auf diese Weise zunächst eine konkrete Einzelentscheidung eines KNN erklären. Darauf aufbauend gibt es jedoch auch Methoden, die Heatmaps konkreter Entscheidungen nutzen, um daraus Muster und damit die abstrakte Funktionsweise des Netzes abzuleiten und eine globale Erklärung eines Modells zu entwerfen.[531]
Einen ähnlichen Ansatz wie *Heatmaps* verfolgen sogenannte *Saliency Maps* („Salienzkarten"), die diejenigen Bildpunkte eines Bildes hervorheben, die maßgeblich für eine Klassifizierung waren.[532]
Einen etwas anderen Herangehensweise haben Techniken, die als *Feature Visualization* („Merkmalsvisualisierung") bezeichnet werden. Sie basieren auf dem Umstand, dass die gelernten Merkmale in einem KNN in unterschiedlichen Bereichen gespeichert werden. Durch *Feature Visualization* kann sichtbar gemacht werden, welche Merkmale einzelne Neuronen, Schichten oder das ganze Netz gelernt haben.[533]
Dies geschieht durch den Entwurf künstlicher Bilder, die den jeweils betrachteten Teil maximal aktivieren und so die für ihn relevanten Merkmale in konzen-

---

[529] Es handelt sich um einen feststehenden Begriff in der XAI-Forschung, weshalb im Folgenden die englische Bezeichnung verwendet wird.

[530] Eine Technik zur Erstellung von Heatmaps ist die sog. Layer-wise Relevance Propagation (LRP), *Bach u. a.*, PLoS ONE 2015, 1; die Methode kann interaktiv hier erprobt werden: *Fraunhofer-Institut für Nachrichtentechnik*, Explainable AI Demos, https://lrpserver.hhi.fraunhofer.de/image-classification (zuletzt abgerufen am 26.10.2023).

[531] So die auf LRP aufbauende Spectral Relevance Analysis (SpRAY), *Lapuschkin u. a.*, Nature Communications 2019, 1.

[532] Eine Übersicht über Veröffentlichungen zu Saliency Maps geben *Guidotti u. a.*, ACM Computing Surveys 2019, 1 (26 ff.). Da sich die Ergebnisse beider Methoden sehr ähneln und es hier nur auf diese Ergebnisse ankommt, wird im Folgenden verallgemeinernd von Heatmaps gesprochen.

[533] Siehe zur Visualisierung von Merkmalen auf der Ebene einzelner Neuronen *Nguyen/Yosinski/Clune*, in: Samek u. a., Explainable AI: Interpreting, Explaining and Visualizing Deep Learning; siehe zur Visualisierung von Merkmalen auf der Ebene von Schichten *Yosinski u. a.*, Deep Learning Workshop, 31st International Conference on Machine Learning 2015, 1.

trierter Form darstellen. Auf diese Weise wird eine Art Prototyp des jeweiligen Merkmals oder der jeweiligen Klasse visualisiert, an dem das Wissen des KNN über bestimmte Merkmale oder ganze Klassen abgelesen werden kann.

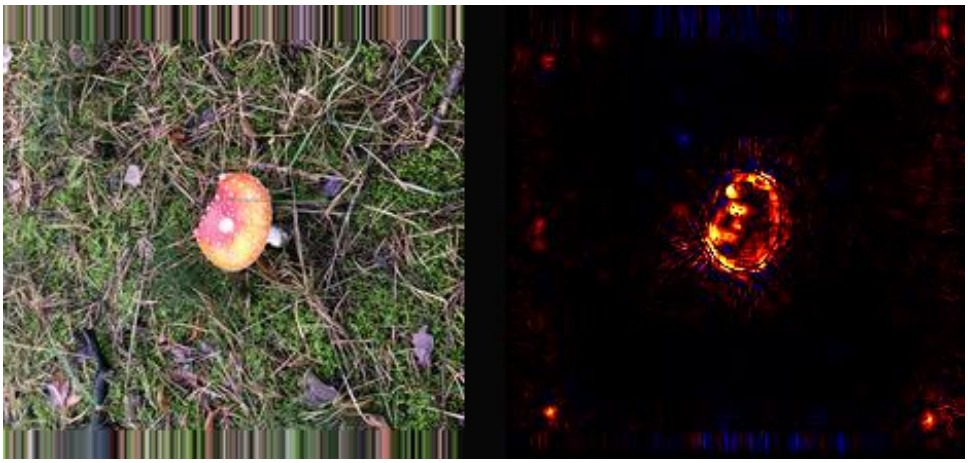

Abb. 7 *Heatmap eines Fliegenpilzes, klassifiziert als „Mushroom" durch die Bildklassifikation (LRP) der Explainable AI Demos des Fraunhofer-Instituts für Nachrichtentechnik.*

Die Beispiele aus dem Bereich der Klassifikation von Bildern zeigen eindrucksvoll, wie der Entscheidungsmechanismus eines KNN verdeutlicht werden kann. Sie sollten jedoch nicht darüber hinwegtäuschen, dass die Logik der Informationsverarbeitung im Netz nicht notwendigerweise der menschlichen Logik entspricht. ML-Techniken sollen ja gerade die Lösung von Aufgaben ermöglichen, die für die klassische Programmierung ungeeignet sind. Ihr Lösungsweg kann daher sogar unintuitiv erscheinen, selbst wenn er durch XAI-Techniken dargestellt werden kann.[534]

Dennoch können *Heatmaps* und *Feature Visualizations* helfen zu verstehen, welche Merkmale für das KNN von Relevanz sind. Die Informationsverarbeitung im Netz wird durch sie symbolisch in einer für den Menschen verständlichen Form repräsentiert und somit explizit. Denn die *Heatmap* verdeutlicht, welche Pixel für die Klassifizierung eines Bildes ausschlaggebend waren. In ihr wird aus der potenziellen semantischen Information des KNN faktische seman-

[534] Siehe zum Ganzen, auch mit visueller Darstellung der Areale, auf die hin einzelne Neuronen einer verdeckten Schicht besonders aktiviert werden, *Burrell*, Big Data & Society 2016, 1 (6 f.).

tische Information. Lediglich das „Warum" der getroffenen Klassifikationsentscheidung bleibt unbeantwortet. Denn *Heatmaps* können zwar zeigen, welche Merkmale relevant für eine Entscheidung waren, aber nicht das Zusammenwirken der verschiedenen Merkmale, ob sie also etwa für sich genommen oder nur im Zusammenspiel relevant sind.[535] Die Beantwortung dieses „Warums" durch den menschlichen Betrachter ist dann auch besonders fehleranfällig.[536]

Dass die Information des KNN in einem *Heatmap* explizit wird, bedeutet jedoch nicht, dass das Netz zugleich reproduzierbar wird. Ganz im Gegenteil: Für die Erstellung der *Heatmap* muss zwar die Aktivierung einzelner Neuronen oder Schichten analysiert werden, diese Analyse wird im Ergebnis jedoch nicht sichtbar. Weder die Architektur des KNN noch seine Gewichte können aus dem *Heatmap* abgelesen werden. Dementsprechend ist es auch nicht möglich, anhand eines *Heatmap* ein funktionsgleiches Netz nachzubauen. Das Geschäftsgeheimnis am KNN wird mithin durch ein *Heatmap* oder eine *Feature Visualization* nicht offenbar.

Darin zeigt sich eine Besonderheit der XAI-Methoden, die sich im weiteren Verlauf der Untersuchung bestätigen wird: Die semantische Information des KNN, also die Regeln der Informationsverarbeitung, wird zwar explizit, sie wird jedoch nicht reproduzierbar offenbart. Und gerade die vor dem Hintergrund der Nacharbeitbarkeit völlig unzureichende Information ist ausreichend, um die implizit repräsentierte Information explizit zu machen. Genauer gesagt ist sie nicht nur ausreichend, sondern sie ermöglicht erst, dass aus potenzieller semantischer Information faktische semantische Information wird.

In den Ergebnissen der XAI-Methode liegen daher Reproduzierbarkeit des KNN sowie Offenbarung des Geschäftsgeheimnisses einerseits und die explizite Darstellung der Information des KNN andererseits weit auseinander. Die Information wird zwar für das menschliche Verständnis zumindest explizit, gleichzeitig ist die Darstellung jedoch so abstrakt, dass eine Reproduktion der dahinterliegenden Informationsverarbeitung nicht möglich ist. Diese Dichotomie ist eine Besonderheit beim Sichtbarmachen der Informationsverarbeitung bei KNN (und anderen Blackbox-Modellen): bei der GOFAI fallen das Vorliegen faktischer semantischer bzw. explizit repräsentierter Information mit deren Re-

[535] *Samek/Müller*, in: Samek u. a., Explainable AI: Interpreting, Explaining and Visualizing Deep Learning, S. 16.

[536] *Samek/Müller*, in: Samek u. a., Explainable AI: Interpreting, Explaining and Visualizing Deep Learning, S. 16.

produzierbarkeit zusammen. Für KNN hingegen müssen sie im Rahmen der Darstellung durch XAI auseinanderfallen.
Diese Diskrepanz kann auch anhand der Unterscheidung der verschiedenen Arten von Opazität verdeutlicht werden. *Heatmaps* vermögen Opazität durch Komplexität zur verringern und können so der Erfüllung von Informationspflichten dienen, die auf die Erklärung einer konkreten Entscheidung zielen. Gleichzeitig beseitigen sie jedoch nicht die Opazität durch Geheimhaltung, denn die eigentliche Information des Netzes, die Regeln der Informationsverarbeitung, werden nicht offenbar.
Dadurch fallen die dargestellten Visualisierungen auch hinsichtlich ihres Erklärungsgehalts in gewisser Weise hinter anderen XAI-Methoden zurück, bei denen die Regeln der Informationsverarbeitung und das Zusammenspiel der Merkmale in größerem Umfang explizit werden.[537]

II. Node-Link-Diagramme

Bei den bereits erwähnten *Node-Link-Diagrammen* können Information und Gewichte farblich oder durch Dicke der Verbindungen dargestellt und anhand ihrer gezeigt werden, wie sich die Eingabedaten durch Aktivierungen der verschiedenen Neuronen und Schichten durch das Netz bewegen.[538] In den letzten Jahren sind viele solcher *Node-Link-Diagramme* als interaktive Modelle umgesetzt werden, anhand derer der Informationsfluss durch ein KNN spielerisch ausprobiert werden kann.[539]
Wie bereits festgestellt, ähnelt die Methode der Darstellung des Netzes als schematischen Graph auf der dritten Darstellungsstufe. Hier zeigt sich, wie schwer die Frage zu beantworten sein kann, ob die Informationsverarbeitung des Netzes explizit wird. Denn werden die Gewichte farblich oder durch Dicke der Verbindungen hervorgehoben, ist das Verständnis der Informationsverarbeitung

[537] Siehe dazu die im Folgenden dargestellten Methoden.

[538] Eine Übersicht über entsprechende Methoden geben bereits *Olden/Jackson*, Ecological Modelling 2002, 135 (139 f.); siehe aus neuerer Zeit auch *Hohman/Kahng/Pienta/Chau*, IEEE Transactions on Visualization and Computer Graphics 2019, 2674 (2683 f.).

[539] Das wohl bekannteste Beispiel ist der „Spielplatz" von TensorFlow: TensorFlow Playground, https://playground.tensorflow.org (zuletzt abgerufen am 26.10.2023).; ein sehr anschauliches Beispiel für den Bereich der Erkennung von handschriftlichen Zahlen gibt *Harley*, An Interactive Node-Link Visualization of Convolutional Neural Networks, https://adamharley.com/nn_vis/mlp/3d.html (zuletzt abgerufen am 26.10.2023).

im Netz, verglichen mit der Darstellung durch Zahlenwerte, schon recht intuitiv. Verstärkt werden kann dieser Effekt durch Visualisierungen, die anzeigen, auf welche Bildbereiche die unterschiedlichen Neuronen oder Schichten eines Netzes „fokussieren".[540] Allerdings zeigen die Beispiele auch, dass die Konzepte, auf die sich die Entscheidung des KNN stützt, nicht notwendigerweise für Menschen verständlichen Konzepten entsprechen. Je nach Art der Visualisierung kann die implizit im Netz gespeicherte, semantische Information in einem *Node-Link-Diagramm* daher schon teilweise explizit werden.
Hinsichtlich der Komplexität gilt jedoch das für die Darstellung als Graph auf Stufe 3 bereits Gesagte: die Darstellung eignet sich gut für verhältnismäßig kleine Netze, ab einer gewissen Größe wird sie unübersichtlich.[541]

III. Diagramme durch LIME und SHAP

KNN sind typische Blackbox-Modelle, deren Regeln der Informationsverarbeitung für den Betrachter unergründlich sind. Es gibt jedoch andere Modelle, die sich aufgrund ihrer Struktur grundsätzlich interpretieren lassen, etwa Entscheidungsbäume oder lineare Modelle, wobei das Verständnis auch hier ab einer gewissen Größe und Komplexität eingeschränkt sein kann. Diese von sich aus erklärbaren Modelle werden auch als *Ante-hoc*-Modelle oder Whitebox-Modelle bezeichnet. Ein gängiger Ansatz zur Erklärung eines Blackbox-Modells ist seine Simulation durch ein interpretierbares Whitebox-Modell, das die Entscheidungen des Blackbox-Modells approximiert. In diesem Fall wird das Ersatzmodell oder Surrogatmodell (engl. *surrogate modell*) also post-hoc eingesetzt, um nachträglich Erklärbarkeit des Blackbox-Modells herzustellen.[542]
Das interpretierbare Ersatzmodell kann dann entweder eine Erklärung für eine konkrete Entscheidung des Blackbox-Modells liefern (sog. *lokale* Erklärbarkeit),

---

[540] Siehe beispielsweise *Harley*, An Interactive Node-Link Visualization of Convolutional Neural Networks, https://adamharley.com/nn_vis/mlp/3d.html (zuletzt abgerufen am 26.10.2023).

[541] So auch *Olden/Jackson*, Ecological Modelling 2002, 135 (140); *Hohman/Kahng/Pienta/Chau*, IEEE Transactions on Visualization and Computer Graphics 2019, 2674 (2684).

[542] Siehe zum Ganzen *Gesellschaft für Informatik*, Technische und rechtliche Betrachtungen algorithmischer Entscheidungsverfahren, S. 55; *Schaaf/Huber*, in: Bitkom, Blick in die Blackbox. Nachvollziehbarkeit von KI-Algorithmen in der Praxis, S. 63; siehe auch Abschnitt A dieses Kapitels.

oder aber es gibt Einblicke in das Modell als Ganzes (sog. *globale* Erklärbarkeit).[543]

Zwei Methoden zur Schaffung lokaler Erklärungen durch Surrogatmodelle werden im Folgenden dargestellt.[544]

Eine weit verbreitete Methode sind die von Ribeiro et al. entwickelten *Local Interpretable Model-agnostic Explanations* (*LIME*).[545] Durch sie kann die durch ein beliebiges Modell getroffene Vorhersage post-hoc erklärt werden, indem lokal für einen bestimmten Vorhersagebereich ein erklärbares Modell entworfen wird.[546] Ein solches erklärbares Modell kann etwa eine lineare Regression oder ein Entscheidungsbaum sein.[547]

Die Methode kann auf ganz unterschiedliche ML-Modelle angewendet werden (modellagnostisch) und mithin auch auf KNN. Es bedarf zudem keines besonderen Wissens über das Blackbox-Modell, der Entwurf von *LIME* kann ohne Kenntnis der Hyperparameter, der Gewichte oder der Aktivierungsfunktion und mithin allein durch Beobachtung des Verhältnisses zwischen Ein- und Ausgabe erfolgen.[548]

Die Ergebnisse des erklärbaren Whitebox-Modells lassen sich dann auf unterschiedliche Weise darstellen. Neben einer Hervorhebung der für eine Klassifikationsentscheidung durch ein KNN relevanten (Super-)Pixel (ähnlich den unter I. beschriebenen Methoden), können etwa die für die inhaltliche Klassifizierung eines Texts relevantesten Wörter in einem Balkendiagramm dargestellt werden.[549] Eine Anwendung ist mithin nicht nur für die Bilderkennung denkbar,

---

[543] *Schaaf/Huber*, in: Bitkom, Blick in die Blackbox. Nachvollziehbarkeit von KI-Algorithmen in der Praxis, S. 63; eine Übersicht von Methoden sowohl zur globalen als auch zur lokalen Erklärbarkeit von Blackbox-Modellen geben *Guidotti u. a.*, ACM Computing Surveys 2019, 1.

[544] Zu globalen Erklärungen sogleich unter 4.

[545] *Ribeiro/Singh/Guestrin*, Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining 2016, 1135.

[546] *Ribeiro/Singh/Guestrin*, Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining 2016, 1135 (1137).

[547] *Dewes*, in: Bitkom, Blick in die Blackbox. Nachvollziehbarkeit von KI-Algorithmen in der Praxis, S. 28.

[548] *Oh/Schiele/Fritz*, in: Samek u. a., Explainable AI: Interpreting, Explaining and Visualizing Deep Learning, S. 124.

[549] So die von Ribeiro et al. gewählten Beispiele, wobei im Fall der Text-Klassifikation kein KNN, sondern eine sog. Support Vector Machine (SVM) als Blackbox-Modell zugrunde lag, *Ribeiro/Singh/Guestrin*, Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining 2016, 1135 (1138).

sondern etwa auch bei ML-Modellen, die für Scoring oder Profiling eingesetzt werden. Dann kann die Bedeutung, die jedes beobachtete Merkmal für eine konkrete Entscheidung hat, graphisch in einem Diagramm dargestellt werden.[550]
Eine andere verbreitete Methode zur lokalen Erklärung der Entscheidungen von Blackbox-Modellen sind die von Lundberg und Lee entwickelten *SHapley Additive exPlanations* (*SHAP*).[551] Sie beruhen auf einer Weiterentwicklung unter anderem von LIME, die mit sog. Shapley-Werten aus der Spieltheorie kombiniert werden.[552] Auch mit der SHAP-Methode wird ein interpretierbares lokales Modell entworfen, mit dessen Hilfe die Entscheidungen des Blackbox-Modells erklärt werden können.[553] Die für einzelne Eingabewerte gewonnen Ergebnisse können dann ebenfalls in unterschiedlichen Diagrammen dargestellt werden. Neben der Darstellung als Balkendiagramm bietet die in Python implementierte SHAP-Bibliothek etwa unter anderem sog. *Beeswarm Plots*, Streudiagramme und Wasserfalldiagramme und sog. *Force Plots* an.[554] Diese verschiedenen Darstellungsmethoden bieten die Möglichkeit, unterschiedliche Aspekte der betrachteten Einzelentscheidung und auch (nicht-lineare) Zusammenhänge zwischen verschiedenen Merkmalen darzustellen. Dazu wird auch eine Einfärbung vorgenommen, anhand derer etwa differenziert werden kann zwischen dem Einfluss, den ein Merkmal auf den Ausgabewert hat („SHAP value“) und der ursprünglichen Gewichtung des Merkmals bei der Eingabe („Feature value“).[555] SHAP kann außerdem auch auf Bilddaten angewendet werden und bietet dann die Möglichkeit der Darstellung eines *Heatmap* zur Verdeutlichung einer Klassifikationsentscheidung.[556]

---

[550] Eine Anwendung von LIME und SHAP auf KNN im Bereich Kredit-Scoring findet sich bei *Misheva u. a.*, 2021, 1.

[551] *Lundberg/Lee*, NIPS 2017, 1.

[552] *Dewes*, in: Bitkom, Blick in die Blackbox. Nachvollziehbarkeit von KI-Algorithmen in der Praxis, S. 28.

[553] *Dewes*, in: Bitkom, Blick in die Blackbox. Nachvollziehbarkeit von KI-Algorithmen in der Praxis, S. 29.

[554] *Lundberg*, SHAP Documentation, https://shap.readthedocs.io/en/latest/api_examples.html#plots (zuletzt abgerufen am 26.10.2023).

[555] So etwa beim Beeswarm Plot: *Lundberg*, SHAP Documentation, https://shap.readthedocs.io/en/latest/example_notebooks/api_examples/plots/beeswarm.html (zuletzt abgerufen am 26.10.2023).

[556] Mit Hilfe des Image Plot, *Lundberg*, SHAP Documentation, https://shap.readthedocs.io/en/latest/example_notebooks/api_examples/plots/image.html (zuletzt abgerufen am 26.10.2023).

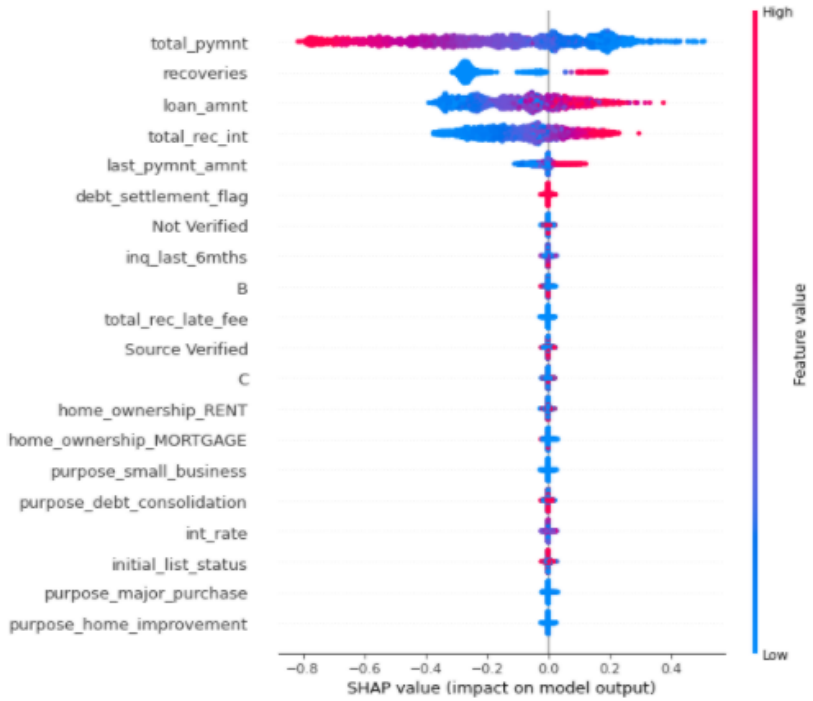

Abb. 8 *Beispiel eines „Beeswarm" Diagramms aus der SHAP Bibliothek, Grundlage ist ein KNN. Quelle: Misheva u. a., arXiv:2103.00949 2021, 1 (12).*

Hinsichtlich der dargestellten Information für den Einsatz in der Bildverarbeitung gilt das zu *Heatmaps* und *Feature Visualizations* Gesagte entsprechend. Interessanter ist die Frage nach dem Informationsgehalt der Diagramme. An diesen kann zunächst abgelesen werden, wie groß der Einfluss jedes einzelnen Merkmals auf die Vorhersageentscheidung des Modells ist. Im Beispiel des *Beeswarm Plots* (**Abb. 8**) hat etwa das Merkmal „total-pymnt", also die bereits erfolgte Tilgung des Kredits, den größten Einfluss auf die Vorhersage. Gleichzeitig zeigt das Beispiel durch die rote Einfärbung, wie stark eine hohe Tilgung die Vorhersage in Richtung einer geringen Ausfallwahrscheinlichkeit ausschlagen lässt.

Die Regeln der Informationsverarbeitung des KNN werden, wie schon bei den *Heatmaps*, in den durch LIME und SHAP geschaffenen Diagrammen nicht offenbar. Dies hat auch einen ganz einfachen Grund: die Diagramme stellen die Informationsverarbeitung der Whitebox-Modelle dar, die als Surrogat des KNN dienen. Dementsprechend wird ihre Annäherung an die Prognoseentscheidun-

gen des KNN dargestellt, die jedoch nicht exakt die Informationsverarbeitung im Netz selbst verrät. Daraus folgt, dass aus der in den Diagrammen offenbarten Information kein funktionsgleiches KNN nachgebaut werden kann. Allenfalls könnte die Information helfen, ein annähernd funktionsgleiches Whitebox-Modell zu entwerfen, also etwa eine lineare oder logistische Regression.[557]
Dennoch kann von einer expliziten Repräsentation der Information des Netzes gesprochen werden. Denn dafür braucht es, wie bereits dargestellt, immer einen Code, anhand dessen die implizit repräsentiere Information symbolisch dargestellt wird. Dieser Code sind im vorliegenden Fall von LIME und SHAP die durch das jeweilige Programm produzierten Diagramme. Dass es für deren Entwurf der Erstellung eines Whitebox-Modells als Zwischenschritt – gewissermaßen als Übersetzung – bedarf, kann im Ergebnis keinen Unterschied machen. Die Ergebnisse der Whitebox-Modelle sind im Vergleich zu denen der Blackbox-Modelle hinreichend genau, um als explizite Darstellung ihrer Information angesehen zu werden.
Dass in den Produkten eines Whitebox-Modells (Diagramme etc.) die Information des KNN explizit wird, bedeutet dennoch nicht, dass mit der Möglichkeit des Nachbaus eines solchen Whitebox-Modells auch das Geschäftsgeheimnis am KNN erlangt ist. Denn die als Geschäftsgeheimnis geschützte Information liegt nicht in den Regeln der Informationsverarbeitung allein – sie liegt eben auch im Subjekt dieser Informationsverarbeitung, dem KNN. Die Annahme, die geheime Information läge im Eingabe-Ausgabe-Verhalten, unabhängig von einem spezifischen Modelltyp, ginge zu weit. Denn die Information über eine in diesem Sinne „abstrakte" Informationsverarbeitung liegt ja auch beim Geheimnisinhaber gar nicht vor. Es kann nur auf die Regeln der Informationsverarbeitung eines trainierten künstlichen neuronalen Netzes ankommen, nicht auf die Regeln der Informationsverarbeitung an sich, losgelöst vom Modell.
Zusammengefasst gilt demnach: die im Netz implizit repräsentierte Information wird durch die Diagramme symbolisch repräsentiert und mithin explizit dargestellt. Gleichzeitig ermöglich sie jedoch nicht die Reproduktion des KNN.

[557] Allerdings sind nach Dewes „Erklärungen [des lokalen Modells] nur für einen sehr begrenzten Wertebereich des Modells gültig", weshalb fraglich ist, ob selbst mit einer Vielzahl von Abfragen ein annähernd funktionsgleiches Modell nachgebaut werden könnte. Vgl. *Dewes*, in: Bitkom, Blick in die Blackbox. Nachvollziehbarkeit von KI-Algorithmen in der Praxis, S. 30.

IV. Entscheidungsbäume

Eine Möglichkeit, globale Erklärungen für die Entscheidungsfindung in einem KNN zu entwerfen, bieten sogenannte Regelextraktionsverfahren. Mit ihrer Hilfe können aus der komplexen Informationsverarbeitung im Netz einfache, für den Menschen verständliche Regeln abgeleitet werden.[558] Diese wiederum lassen sich dann auf unterschiedliche Weise darstellen, etwa als Wenn-dann-Regel oder in der Form eines Entscheidungsbaums.

Wie gezeigt sind die Regeln der Informationsverarbeitung in tiefen neuronalen Netzen derart komplex, dass eine Ableitung verständlicher Regeln jedoch alles andere als banal ist und die Forschung schon lange beschäftigt.[559] Abgeleitete Entscheidungsbäume werden schnell zu groß oder die Übereinstimmung mit den Prognosen des KNN ist ungenügend. Die Regelextraktion kann jedoch vereinfacht werden, indem sie bereits beim Trainieren des Blackbox-Modells durch bestimmte Anpassungen vorbereitet wird, sodass das trainierte Modell sich inhärent besser für die Ableitung eines Entscheidungsbaums eignet. Auf diese Weise können für tiefe, jedoch verhältnismäßig einfach aufgebaute KNN verständliche Entscheidungsbäume mit guter Genauigkeit entworfen werden.[560]

Das Ergebnis kann zur Erfüllung einer Transparenzpflicht durchaus geeignet sein, denn Entscheidungsbäume gelten als besonders gut interpretierbar und leicht verständlich.[561] Mithilfe eines durch ein Surrogatmodell erstellten Entscheidungsbaums kann abstrakt erklärt werden, welche Merkmale die Entscheidung eines Modells beeinflussen werden – diese Möglichkeit globaler Erklärungen qualifiziert sie für den Einsatz zur Erfüllung von ante-hoc Transparenzpflichten.

Im Entscheidungsbaum wird die Information des KNN auch symbolisch repräsentiert und mithin explizit. Es handelt sich um einen von den soeben dargestell-

---

[558] Siehe zur Regelextraktion beim Kredit-Scoring *Thomas/Edelman/Crook*, Credit scoring and its applications, S. 70; siehe auch das patentierte Modell NeuroDecision von Equifax.

[559] Eine Übersicht entsprechender Techniken geben bereits 1995 *Andrews/Diederich/Tickle*, Knowledge-Based Systems 1995, 373; Craven entwickelt bereits 1996 den Algorithmus TREPAN zur Extraktion von Entscheidungsbäumen aus KNN. Siehe mit Illustrationen aus unterschiedlichen Anwendungsfeldern: *Craven*, Extracting comprehensible models from trained neural networks.

[560] Siehe zur Ableitung eines Entscheidungsbaums aus einem derart „vorbereiteten" Multilayer Perceptron *Schaaf/Huber*, in: Bitkom, Blick in die Blackbox. Nachvollziehbarkeit von KI-Algorithmen in der Praxis, S. 64 ff.

[561] *Guidotti u. a.*, ACM Computing Surveys 2019, 1 (16).

ten Diagrammen verschiedenen Code, der – im Unterschied zu diesen – auch einen Entscheidungsweg aufzeigt. Denn anders als die Diagramme kann an einem Entscheidungsbaum nicht nur das Ergebnis einer Klassifikationsentscheidung, sondern eben auch der Lösungsweg abgelesen werden. Da die Entscheidungsfindung im Netz und mithin die Regeln seiner Informationsverarbeitung durch den Entscheidungsbaum angenähert werden, kann bis zu einem gewissen Grad nachvollzogen werden, welche Kombinationen von Merkmalen beziehungsweise von Merkmalsgewichtungen zu welchem Ergebnis führen.
Über die tatsächliche Informationsverarbeitung im Netz sagt der Entscheidungsbaum, dem ebenfalls die „Zwischenübersetzung" des Surrogatmodells zugrunde liegt, jedoch ebenso wenig aus wie die Diagramme. Weder die Netzarchitektur noch die Gewichte lassen sich aus dem Entscheidungsbaum herauslesen und dementsprechend kann auch kein funktionsgleiches Netz nachgebaut werden. Auch hier fallen mithin explizite Darstellung der Information des KNN und Reproduzierbarkeit auseinander.

V. Kontrafaktische Erklärungen

Eine weitere Möglichkeit zur Erklärung einer konkreten Entscheidung eines KNN besteht darin zu verdeutlichen, wie eine veränderte Eingabe die Ausgabe des KNN beeinflussen würde.
Diese sogenannten kontrafaktischen Erklärungen (*counterfactual explanations*) bieten mit recht einfachen Mitteln die Chance, eine nachvollziehbare Erklärung der Entscheidung eines Modells zu geben.[562] Generiert werden können sie durch sog. *Adversarial Perturbations* (wörtlich: „feindliche Störungen"), bei denen ein synthetischer Datenpunkt geschaffen wird, der sehr nahe am ursprünglichen Datenpunkt liegt, für den das Netz dann jedoch eine andere Klassifikationsentscheidung trifft.[563]
Eine Erklärung könnte dann nach Wachter et al. so aussehen:

---

[562] *Dewes*, in: Bitkom, Blick in die Blackbox. Nachvollziehbarkeit von KI-Algorithmen in der Praxis, S. 23; ein Beispiel ist die Methode „LOcal Rule-based Explanations" (LORE), welche die Regelextraktion für eine konkrete Entscheidung (lokal) ermöglicht und dann auch Gegenbeispiele gibt, bei welchen Merkmalsveränderungen die Entscheidung anders ausgefallen wäre. Siehe dazu *Guidotti u. a.*, IEEE Intelligent Systems 2019, 14.

[563] *Wachter/Mittelstadt/Russell*, Harvard Journal of Law & Technology 2018, 841 (852); *Dewes*, in: Bitkom, Blick in die Blackbox. Nachvollziehbarkeit von KI-Algorithmen in der Praxis, S. 23.

> „Score p was returned because variables V had values (v_1, v_2, . . .) associated with them. If V instead had values (v_1', v_2', . . .), and all other variables had remained constant, score p' would have been returned."[564]

Grundlegender Gedanke hinter kontrafaktischen Erklärungen ist es, die Eingabe nur so viel zu ändern wie unbedingt nötig, um eine signifikante Änderung der Aussage zu erhalten – im Falle eines Klassifikationsmodells also etwa die Einordnung in eine andere Klasse.[565]
Je nach Fragestellung können auch kontrafaktische Erklärungen hinsichtlich verschiedener Variablen sinnvoll sein, um eine Modellentscheidung stichhaltig zu erklären und unterschiedliche Handlungsmöglichkeiten für die betroffene Person aufzuzeigen.[566]
Aufgrund ihres äußerst eingeschränkten Informationsgehalts ist in kontrafaktischen Erklärungen nur ein winziger Teil der semantischen Information des KNN enthalten. Dieser kleine Teil wird in der Erklärung einer einzelnen oder mehrerer Entscheidungen durch Alternativbeispiele zwar explizit. Dennoch ermöglicht diese explizite Teilinformation weder einen Nachbau des KNN noch sind Konstellationen denkbar, in denen sie im Zusammenspiel mit weiteren Informationen einen solchen Nachbau wesentlich begünstigen könnte.
Auch in kontrafaktischen Erklärungen werden mithin Teile der Information eines KNN explizit, ohne dass dadurch die Reproduktion des Netzes ermöglicht würde.

VI. Weitere Methoden

Es gibt unzählige weitere spannende Ansätze im Bereich der XAI. So können die Entscheidung eines KNN zur Bilderkennung etwa in Textform[567] oder in

---

[564] *Wachter/Mittelstadt/Russell*, Harvard Journal of Law & Technology 2018, 841 (848).

[565] Wachter u.a. sprechen vom Konzept der „closest possible world", *Wachter/Mittelstadt/Russell*, Harvard Journal of Law & Technology 2018, 841 (845); siehe auch *Dewes*, in: Bitkom, Blick in die Blackbox. Nachvollziehbarkeit von KI-Algorithmen in der Praxis, S. 23.

[566] *Wachter/Mittelstadt/Russell*, Harvard Journal of Law & Technology 2018, 841 (851).

[567] Eine von Hendricks u.a. entwickelte Methode begründet die Klassifikation von Vögeln in Textform, z. B.: „This is a Western Grebe because this bird has a long white neck, pointy yellow beak and red eye.", *Hendricks u. a.*, in: Leibe/Matas/Sebe/Welling, Computer Vision – ECCV 2016, S. 2.

gesprochener Sprache begründet werden.[568] Während diese unterschiedlichen Methoden aus Sicht des Verbraucherschutzes und der Regulierung von gesteigertem Interesse sein können, bieten sie aus der Sicht des Geheimnisschutzes und mithin für die vorliegende Arbeit keinen großen Erkenntnisgewinn. Denn die Erklärung der Entscheidung eines KNN in natürlicher Sprache[569] ist weniger präzise als die Diagramm-Modelle und bietet einen ähnlichen Informationsgehalt wie die soeben dargestellten kontrafaktischen Erklärungen.

C. Transparenzpflichten und Geheimnisschutz

I. Heatmaps und Feature Visualizations

Heatmaps und *Feature Visualizations* machen einerseits die Information des KNN explizit, erlauben aber andererseits nicht dessen Reproduktion. Indem sie mithin Opazität durch Komplexität, nicht jedoch Opazität durch Geheimhaltung beseitigen, eignen sie sich besonders gut für die Erfüllung von Transparenzpflichten gegenüber Laien.

1. Datenschutzgrundverordnung

Beide Methoden könnten eine gute Möglichkeit darstellen, den Informations- und Auskunftsrechten hinsichtlich der „involvierten Logik" des Modells gemäß Art. 13 Abs. 2 lit. f, Art. 14 Abs. 2 lit. g und Art. 15 Abs. 1 lit. h DSGVO zu begegnen. Durch sie wird einerseits das Erfordernis der Verständlichkeit gewahrt, andererseits wird die Offenbarung von Geschäftsgeheimnissen vermieden und eine, wie auch immer gelagerte, Abwägungsentscheidung erübrigt sich. Einschränkend gilt lediglich, dass es sich typischerweise um ex-post-Methoden handelt, die lediglich eine konkrete Entscheidung erklären können. Für die (ex-ante) Informationspflichten der Art. 13 und 14 DSGVO könnten daher allenfalls repräsentative Beispiele oder eine interaktive Methode dienen, um die abstrakte Funktionsweise des Modells zu verdeutlichen. Denn auch aus der Beobachtung

[568] *Ehsan/Riedl*, in: Stephanidis/Kurosu/Degen/Reinerman-Jones, HCI International 2020 - Late Breaking Papers: Multimodality and Intelligence.

[569] Nicht zu verwechseln mit der Erklärung des Aufbaus des KNN in natürlicher Sprache, wie in Kapitel 7 beschrieben.

eines Modells in Bezug auf viele Einzelentscheidungen dürfte zumindest teilweise auf die Funktionsweise des Modells geschlossen werden können.[570]

Durch *Heatmaps* und *Feature Visualizations* können sowohl die wichtigsten Parameter als auch deren ungefähre Gewichtung abgelesen werden. Allerdings erscheinen beide Visualisierungsmethoden für die klassischerweise im Rahmen von Art. 13 bis 15 DSGVO diskutierten Anwendungsfälle weniger geeignet, da es sich in den meisten Fällen nicht um die Klassifikation von Bildmaterial handeln wird. Teilweise wird der Einsatz von ähnlichen Visualisierungstechniken jedoch auch für die Erfüllung von Transparenzpflichten hinsichtlich hochkomplexer Modelle in anderen Bereichen als der Bilderkennung erwogen.[571]

2. Entwurf der KI-Verordnung

a) Überblick

Für die Erfüllung der sich aus dem KI-VO-E ergebenden Transparenzpflichten sind *Heatmaps* und *Feature Visualizations* nur teilweise geeignet.
Vorweggeschickt sei zunächst, dass im Anwendungsbereich des KI-VO-E Verfahren der Bilderkennung eine größere Rolle spielen dürften als im Anwendungsbereich der Art. 13 bis 15 DSGVO. Und auch wenn die Verständlichkeit im Rahmen des KI-VO-E eine wesentlich kleinere Rolle spielt als in der DSGVO, kommen XAI-Methoden auch hier zur Erfüllung der Transparenzpflichten grundsätzlich in Frage.
Ein relevanter Bereich, in dem KNN zur Bilderkennung Anwendung finden, ist die medizinische Diagnostik. Entsprechende Software dürfte auch als Medizinprodukt einzuordnen sein im Sinne des Art. 2 Nr. 1 EU-Medizinprodukte-Verordnung[572] und daher gemäß Art. 6 Abs. 1 lit. a KI-VO-E in Verbindung mit Abschnitt A Nr. 11 des Anhang II KI-VO-E auch ein KI-Hochrisiko-System

---

[570] Samek und Müller sprechen in diesem Zusammenhang von „Meta-explanations", *Samek/Müller*, in: Samek u. a., Explainable AI: Interpreting, Explaining and Visualizing Deep Learning, S. 14.

[571] Koch/Biggen sehen die Möglichkeiten, auf diese Weise Transparenzverpflichtungen aus Art. 5 P2B-VO nachzukommen, *Koch/Biggen*, NJW 2020, 2921 (2925).

[572] Verordnung (EU) 2017/745 des Europäischen Parlaments und des Rates vom 5. April 2017 über Medizinprodukte, zur Änderung der Richtlinie 2001/83/EG, der Verordnung (EG) Nr. 178/2002 und der Verordnung (EG) Nr. 1223/2009 und zur Aufhebung der Richtlinien 90/385/EWG und 93/42/EWG des Rates (ABl. L 117 vom 5.5.2017, S. 1)

sein, für das unterschiedliche Transparenzpflichten des KI-VO-E gelten. Hier könnten *Heatmaps* etwa helfen, Ärzten im Rahmen der in Art. 13 KI-VO-E geforderten Transparenz die Entscheidung eines KNN nachvollziehbar zu machen, ohne gleichzeitig die Offenbarung des Geschäftsgeheimnisses durch Reproduzierbarkeit des Modells zu bewirken.
Auch zur Erklärung der Entscheidungen von Hochrisiko-KI-Systemen zur biometrischen Identifizierung und Kategorisierung natürlicher Personen gemäß Art. 6 Abs. 3 KI-VO-E in Verbindung mit Nr. 1 des Anhang III KI-VO-E könnten *Heatmaps* und *Feature Visualizations* geeignet sein.
Einschränkungen könnten sich zwar daraus ergeben, dass die Transparenzpflichten nach dem Sinn und Zweck des KI-VO-E nur auf eine Erklärung der abstrakten Funktionsweise eines KNN und nicht lediglich einer konkreten Entscheidung abzielen können. Denn die zuständigen nationalen Behörden und die notifizierten Behörden sollen die Konformität des Systems – und eben nicht einer konkreten Entscheidung – mit dem Verordnungsentwurf und anderen Rechtsvorschriften überprüfen können. Und auch der Nutzer soll „die Ergebnisse des Systems angemessen verstehen und verwenden können". Bei diesen Ergebnissen handelt es sich zwar um konkrete Entscheidungen, allerdings eben auch um eine Vielzahl, weshalb ein *Heatmap* einer einzelnen Entscheidungsfindung dem Nutzer bei der Interpretation des Modells kaum weiterhelfen wird. *Heatmaps* als typische Post-hoc-Methoden, die keine Erklärung der abstrakten Funktionsweise eines Modells zulassen, scheinen daher für die anvisierten Erklärungen in der technischen Dokumentation oder in einer Gebrauchsanweisung auf den ersten Blick kaum geeignet. Allenfalls eine große Zahl von *Heatmaps* konkreter Entscheidungen könnte Rückschlüsse über die abstrakte Funktionsweise des Netzes ermöglichen, deren Abbildung in den entsprechenden Unterlagen scheint jedoch kaum praktikabel.
Allerdings dürfte im Regelungsbereich des KI-VO-E zu überlegen sein, ob die klassische Vorstellung einer technischen Dokumentation und einer Gebrauchsanweisung als Schriftstücke nicht überholt wäre. Dafür spricht, dass Hochrisiko-KI-Systeme nach Art. 13 Abs. 2 KI-VO-E mit Gebrauchsanweisungen „in einem geeigneten digitalen Format bereitgestellt oder auf andere Weise mit Gebrauchsanweisungen versehen" werden sollen. Es erscheint daher durchaus denkbar, den sich aus dem KI-VO-E ergebenden Transparenzverpflichtungen auch durch interaktive Darstellungsmethoden nachzukommen, die entweder

als Programm mitgeliefert werden oder auf die online (für Behörden und notifizierte Stellen auch über Schnittstellen) zugegriffen werden könnte.
Im Einzelnen ist der Einsatz von *Heatmaps* und *Feature Visualizations* im Rahmen des KI-VO-E wie folgt zu bewerten.

b) Private Informationsempfänger

Zur Erfüllung der Transparenzpflicht des Art. 13 KI-VO-E gegenüber dem Nutzer eines KNN scheinen *Heatmaps* und *Feature Visualizations* gut geeignet. Sie dürften dem Nutzer das „angemessene" Verstehen und Verwenden der Ergebnisse des Systems und die Erfüllung seinen Pflichten aus dem KI-VO-E ermöglichen (Art. 13 Abs. 1 KI-VO-E). Ob sie eine „präzise, vollständige, korrekte und eindeutige Informationen in einer für die Nutzer relevanten, barrierefrei zugänglichen und verständlichen Form" im Sinne des Art. 13 Abs. 2 KI-VO-E darstellen, kann zwar hinsichtlich Präzision und Vollständigkeit bezweifelt werden. Allerdings scheint die Erfüllung dieser Anforderungen in der Gesamtheit für hochkomplexe KNN aufgrund der aufgezeigten Opazität durch Komplexität unmöglich.
Es wird daher von der Auslegung der unbestimmten Rechtsbegriffe in Art. 13 Abs. 1 bis 3 KI-VO-E abhängen, ob die Darstellung durch *Heatmaps* und *Feature Visualizations* letztlich für ausreichend erachtet wird. Da sich Parameter und ihr Einfluss auf die Klassifikationsentscheidung an beiden Visualisierungen ungefähr ablesen lassen, weiß der Nutzer zumindest ungefähr, worauf die Entscheidung des Modells beruht („angemessen verstehen") und kann daraufhin entscheiden, ob er das Ergebnis für seine weitere Arbeit verwendet. Auch grobe Fehler in der Klassifikation dürften mit beiden Techniken sichtbar werden.[573] Die Darstellung hätte den großen Vorteil, dass die Gewichte als sensibelste Information des Geheimnisinhabers nicht offenbart werden müssten. Sie scheint daher eine gute Möglichkeit, einen angemessenen Ausgleich zwischen Transparenz und dem Schutz von Geschäftsgeheimnissen zu schaffen.

c) Öffentliche Informationsempfänger

Für die Transparenzverpflichtungen gegenüber notifizierten Behörden dürften beide XAI-Methoden ebenfalls grundsätzlich geeignet sein. Notifizierte Behör-

[573] Siehe zum interaktiven Einsatz von *Heatmaps* und *Feature Visualizations* in diesem Kapitel B 1.

den stützen ihre Konformitätsbewertung unter anderem auf die technische Dokumentation (vgl. Anhang VII Nr. 1 KI-VO-E). Ob die darin anzugebenden Bedeutung der verschiedenen Parameter für die Optimierung des Systems auch die Gewichte eines KNN umfasst, lässt sich nicht eindeutig beantworten.[574] Aus Sicht des Geheimnisschutzes wäre eine restriktive und mithin verneinende Auslegung zu begrüßen, da jede Offenlegung das Risiko des faktischen Geheimnisverlustes birgt. Die entscheidende Frage ist daher, ob sich mit Techniken wie *Heatmaps* oder *Feature Visualizations* die Bedeutung der Parameter für die Entscheidung eines KNN in einer für die technische Dokumentation und deren Bewertung durch notifizierte Stellen ausreichenden Genauigkeit darstellen lassen können. Grundsätzlich ermöglichen beide Techniken, den Einfluss einzelner Bildregionen beziehungsweise einzelner Merkmale auf die Entscheidung eines KNN darzustellen. Prinzipiell erscheint es daher möglich, die „Bedeutung der Parameter" einer Entscheidung auch anhand dieser Methoden abzulesen und das System etwa anhand interaktiver *Heatmaps* zu testen. Dies hätte den großen Vorteil, dass die technische Dokumentation, die neben den notifizierten Stellen auch weiteren zuständigen Behörden und öffentlichen Stellen zur Verfügung zu stellen sein dürfte, nicht die sensiblen Gewichte eines KNN enthalten müsste. Sollte die Darstellung mittels XAI-Methoden im Einzelfall zu ungenau sein, bestünde immer noch die Möglichkeit, die Gewichte auf begründeten Antrag zur Verfügung zu stellen analog der in Nummer 4.5. des Anhang VII KI-VO-E vorgesehenen Möglichkeit für den Quellcode. Diese dürfte für KNN so auszulegen sein, dass auch der Zugang zu den Gewichten und Schwellenwerten gewährt werden müsste, da anhand des Quellcodes allein die Funktionalität eines KNN nicht überprüft werden kann.

Für die Offenlegungspflichten gegenüber der Marktüberwachungsbehörde gemäß Art. 23 und Art. 63 Abs. 8 und 9 KI-VO-E gelten die obigen Ausführungen entsprechend. Nur wird ihr gegenüber die Offenlegung der Gewichte noch am ehesten erforderlich und auch zu rechtfertigen sein, da die Marktüberwachungsbehörde am umfangreichsten zur Prüfung ermächtigt wird.

Für Transparenzpflichten gegenüber den übrigen nationalen Behörden und insbesondere gegenüber den in Art. 64 KI-VO-E genannten öffentlichen Stellen gilt die zu notifizierten Stellen vorgenommene Bewertung erst recht. Für ihre Zwecke dürfte die Darstellung der Entscheidungen eines Hochrisiko-KI-Systems anhand eines *Heatmaps* in der technischen Dokumentation ausreichend

[574] Siehe dazu bereits oben, Kapitel 3 C II.

sein. Auch für sie besteht nämlich in Art. 64 Abs. 5 S. 1 KI-VO-E die Möglichkeit, bei der Marktüberwachungsbehörde einen begründeten Antrag auf Durchführung technischer Tests des Hochrisiko-KI-Systems stellen.

II. Node-Link-Diagramm

Inwieweit in einem *Node-Link-Diagramm* die semantische Information des KNN faktisch wird und sich die Methode daher für die Erfüllung von Transparenzverpflichtungen eignet, ist eine Frage des Einzelfalls. Maßgebliche Faktoren sind die Komplexität des Netzes und der Umstand, ob es sich bei den in verborgenen Schichten stattfindenden Verarbeitungen um für den Menschen nachvollziehbare Konzepte handelt, was meist nicht der Fall sein wird.
Gerade im Bereich des *Deep Learning*, in dem der Einsatz von KNN besonders relevant ist, wird daher häufig ebenso wenig die Entstehung faktischer semantischer Information anzunehmen sein wie bei der Darstellung als Graph. Denn auch wenn Variationen in Stärke und Farbe der Verbindungen eine intuitivere Vorstellung von der Funktionsweise eines KNN vermitteln können als die bloße Darstellung als Graph mit Katengewichten, so bleibt die tatsächliche Informationsverarbeitung und Entscheidungsfindung des Netzes doch auf ähnliche Weise im Dunkeln. Auch interaktive Möglichkeiten[575] vermögen zwar anschaulich darzustellen, wie die unterschiedlichen Neuronen und Schichten zusammenwirken. Eine belastbare Erklärung, auf welche Merkmale eine Entscheidung in welchem Maße gestützt wird, ist dieser Art von Darstellungen allerdings nicht zu entnehmen. Die Opazität durch Komplexität wird kaum verringert. Die Bewertung unterscheidet sich in diesem Punkt mithin kaum von der zur Darstellung als Graph auf der dritten Stufe.
Anders verhält es sich bei der Frage der Reproduzierbarkeit. Dem *Node-Link-Diagramm* lassen sich, wie der Darstellung als Graph, die Anzahl der Neuronen und Schichten und mithin die Architektur des Netzes entnehmen. Dabei handelt es sich bereits um einen sensiblen Teil der Information des KNN, die das Geschäftsgeheimnis ausmacht. Die Opazität durch Geheimhaltung wird mithin zu einem gewissen Grad beseitigt. Anders als bei der Darstellung als Graph mit gewichteten Verbindungen lassen sich dem *Node-Link-Diagramm* jedoch nicht die Gewichte des KNN entnehmen. Allein die Stärke der Verbindungen und Hinweise auf positive oder negative Aktivierung lassen diesbezüglich keine exak-

[575] Siehe etwa zum *Tensorflow Playground* oben, Fn. 539.

ten Schlüsse zu. Der wertvollste Teil der Information des KNN wird mithin in dieser Darstellungsform nicht offenbar.
Ob *Node-Link-Diagramme* sich für die Erfüllung von Transparenzpflichten eignen, lässt sich daher nicht allgemein beantworten. Sie werden einerseits in vielen Fällen aufgrund der Größe und Komplexität des Netzes keine hilfreichen Erklärungen für die Entscheidung eines KNN liefern, andererseits sind sie aus Geheimnisschutzgründen problematischer als andere XAI-Methoden.

1. Datenschutzgrundverordnung

Bei kleineren Netzen könnte sich ein *Node-Link-Diagra*mm für die Darstellung der involvierten Logik einer automatisierten Entscheidungsfindung im Sinne der Art. 13 Abs. 2 f, Art. 14 Abs. 2 g und Art. 15 Abs. 1 h DSGVO grundsätzlich eignen. Dies gilt umso mehr, sollten die einzelnen Schichten oder Neuronen verständliche Konzepte abbilden. Zumindest sofern die Informations- und Auskunftspflicht restriktiv nur als auf die Parameter und deren Gewichtungen beschränkt ausgelegt wird, könnten *Node-Link-Diagramme* eine gute Erklärung liefern. Denn sowohl die Eingabeparameter als auch ihre Relevanz für die getroffene Entscheidung lassen sich an einem kleinen *Node-Link-Diagramm* näherungsweise ablesen.
Eine Einschränkung besteht jedoch dahingehend, dass es sich um eine lokale Methode handelt, die mithin grundsätzlich keine Erklärung der abstrakten Funktionsweise des Modells gibt. Den (ex-ante) Informationspflichten der Art. 13 und 14 DSGVO könnte daher nur insofern nachgekommen werden, als ein oder mehrere *Node-Link-Diagramme* beispielhaft zur Erklärung des Systems zur Verfügung gestellt würden. Auch eine interaktive Methode wäre für diesen Fall denkbar.
Allerdings wird durch die Darstellung Information über die Netzarchitektur offenbar, die als Geschäftsgeheimnis geschützt sein kann, und die in Kombination mit weiterer Information zu einer Offenbarung des Geschäftsgeheimnisses an einem KNN führen könnte. Andere XAI-Methoden, die keine Information über die Netzarchitektur offenbaren, dürften daher im Bereich der DSGVO sinnvoller sein.[576]

[576] Siehe etwa die nachfolgend dargestellten Diagramme durch LIME und SHAP.

2. Entwurf der KI-Verordnung

Dem Nutzer im Sinne des KI-VO-E in der Gebrauchsanweisung ein *Node-Link-Diagramm* zur Verfügung zu stellen, erscheint aus Sicht des Geheimnisschutzes vertretbar, da die Netzarchitektur dem Nutzer ohnehin im Rahmen der Information über die „Merkmale" des Systems zur Verfügung zu stellen sein dürfte (vgl. Art. 13 Abs. 3 lit. b KI-VO-E). Ein (interaktives) *Node-Link-Diagramms* könnte dem Nutzer die Funktionsweise des KI-Systems in „verständliche[r] Form" (Art. 13 Abs. 2 KI-VO-E) verdeutlichen und ihn befähigen, das System „angemessen [zu] verstehen und [zu] verwenden" im Sinne des Art. 13 Abs. 1 KI-VO-E.

Die Frage, ob sich *Node-Link-Diagramme* für die Erfüllung der gegenüber zuständigen nationalen Behörden, anderen öffentlichen Stellen und notifizierten Stellen im KI-VO-E vorgesehenen Transparenzpflichten eignen, muss differenziert beantwortet werden. Dabei kommt es maßgeblich darauf an, ob die Offenlegung der Gewichte für erforderlich gehalten wird, was sich aus dem Wortlaut der relevanten Vorschriften nicht eindeutig ermitteln lässt.[577]

Im Sinne eines effektiven Geschäftsgeheimnisschutzes schiene es zumindest nicht ausgeschlossen, die Offenlegung der Gewichte an die gestaffelten Regeln zur Offenlegung des Quellcodes zu koppeln, wie sie im KI-VO-E vorgesehen sind.

Dann könnten auch *Node-Link-Diagramme* eine geeignete Methode zur Darstellung der Bedeutung der Parameter in der technischen Dokumentation sein. Zwar lässt sich an ihnen, anders als an *Heatmaps* und *Feature Visualizations*, die Architektur des Netzes ablesen. Sie sind daher aus Sicht des Geheimnisschutzes kritischer als zuerst untersuchten Visualisierungstechniken. Allerdings dürfte auch hier die Architektur des Netzes ohnehin im Rahmen der technischen Dokumentation zu offenbaren sein.[578] Wie die bereits im vorangegangenen Abschnitt untersuchten XAI-Methoden scheinen *Node-Link-Diagramme* für die Darstellung der Bedeutung der verschiedenen Parameter im Sinne der Nr. 2b Anhang IV KI-VO-E durchaus geeignet.

---

[577] Siehe dazu oben, Kapitel 3 C II.

[578] Auch wenn mit „Systemarchitektur" im Sinne von Nr. 2c des Anhang IV KI-VO-E wohl nicht die Architektur des Netzes, sondern die Architektur der unterschiedlichen Softwarekomponenten gemeint sein dürfte, so wird die Netzarchitektur unter die „allgemeine Logik des Systems" im Sinne der Nr. 2b des Anhang IV KI-VO-E zu subsumieren sein.

III. Diagramme durch LIME und SHAP

Das obige Beispiel einer SHAP-Anwendung aus dem besonders grundrechtsrelevanten Bereich des Kreditscorings zeigt eindrücklich, wie gut die visuelle Darstellung der Vorhersageentscheidung eines KNN in Diagrammen für die Erfüllung von Transparenzpflichten gegenüber Laien geeignet ist. Es lässt sich intuitiv erfassen, welche Merkmale für die Kreditentscheidung relevant waren und auch, in welchem Maß ihr jeweiliger Wert den Ausgang beeinflusst hat. Gleichzeitig ist der Offenbarungsgehalt mit Blick auf die Information des KNN, also die Regeln der Informationsverarbeitung, gering. Sofern das Diagramm alle Merkmale (und nicht nur etwa die wichtigsten zehn) auflistet, lässt sich zumindest die Größe von Eingabe- und Ausgabeschicht mit ziemlicher Sicherheit bestimmen. Auch wird offenbar, auf welche Merkmale die Entscheidung gestützt wird.

Darüberhinausgehende Informationen über die Architektur des Netzes lassen sich den Diagrammen nicht entnehmen. Und auch die Gewichtungen der Merkmale haben keinerlei Aussagekraft über die Gewichte des KNN, die sich auf die Verbindungen des gesamten Netzes verteilen. Die durch LIME- oder SHAP-Diagramme offenbarte Information hat sich somit gewissermaßen zu weit von der Information des Netzes entfernt, sie ist zu abstrakt geworden als dass sie das Geschäftsgeheimnis am KNN gefährden könnte.

Für die Erfüllung der hier untersuchten Transparenzpflichten ergibt sich daher ein ähnliches Bild wie bei den bereits dargestellten XAI-Methoden.

1. Datenschutzgrundverordnung

Sofern im Rahmen der Art. 13 Abs. 2 f, Art. 14 Abs. 2 g und Art. 15 Abs. 1 h DSGVO nur eine Offenlegung von Parametern und Gewichtung und nicht des „Algorithmus“ gefordert wird, bieten LIME und SHAP gute Methoden zur Erfüllung dieser Verpflichtung. Der Betroffene kann an den Diagrammen mit großer Genauigkeit ablesen, welche Merkmale welchen Einfluss auf die automatisierte Entscheidung hatten, ohne dass Geschäftsgeheimnisse offenbart würden.[579] Sowohl hinsichtlich des Parameters der Verständlichkeit (Art. 12 Abs. 1

[579] Die bloße Information über wesentliche Parameter wird hier nicht als kritisch für den Schutz des Geschäftsgeheimnisses erachtet, da sie in den meisten Fällen in Fachkreisen bekannt sein wird. Anders insofern Hacker, der die Offenlegung der „wichtigsten Entscheidungsfaktoren“ nur bei „besonders risikoträchtigen Apllikationen“ fordert, *Hacker*, NJW 2020, 2142 (2144).

DSGVO) als auch hinsichtlich des Geheimnisschutzes sind die Methoden mithin sehr gut geeignet. Hinsichtlich der Erklärung der abstrakten Funktionsweise des Modells (Art. 13 und 14 DSGVO) gilt das bereits zu den vorangehenden XAI-Methoden Gesagte entsprechend. Beispiele und interaktive Methoden dürften auch eine Erklärung des Modells ermöglichen, bevor überhaupt eine konkrete Entscheidung getroffen wird.

2. Entwurf der KI-Verordnung

Diagramme aus LIME und SHAP dürften auch für die Gebrauchsanweisung, die dem Nutzer im Sinne des KI-VO-E zur Verfügung gestellt werden soll, gut geeignet sein. Dies gilt zumindest dann, wenn sie in interaktiver Form angeboten werden und somit die Überprüfung vieler Einzelentscheidung ermöglichen. Der Nutzer könnte mit ihrer Hilfe die wesentlichen Parameter und ihre Bedeutung für das Ergebnis des Systems „angemessen verstehen und verwenden" im Sinne des Art. 13 Abs. 1 KI-VO-E. Die Information wäre „präzise" und „verständlich" und könnte – je nach Umfang des Diagramms – wohl auch als „vollständig" im Sinne des Art. 13 Abs. 2 KI-VO-E bezeichnet werden. Für den Geheimnisinhaber wären die Darstellungsmethoden von Vorteil, da sie kaum geheime Information über das KNN preisgeben. Sofern auch Information über die Netzarchitektur im Rahmen des Art. 13 Abs. 3 lit. b KI-VO-E zur Verfügung gestellt werden müsste, die sich aus den Diagrammen nicht herauslesen lässt, könnte dies separat erfolgen.

Auch für die Frage, ob sich mithilfe von LIME und SHAP hergestellte Diagramme für die Erfüllung der gegenüber zuständigen nationalen Behörden, anderen öffentlichen Stellen und notifizierten Stellen im KI-VO-E vorgesehenen Transparenzpflichten eignen, kommt es maßgeblich darauf an, ob die Offenlegung der Gewichte für erforderlich gehalten wird.[580]

Sollte dies verneint werden, so könnten auch LIME und SHAP geeignete Methoden sein, um die Bedeutung der Parameter in der technischen Dokumentation darzustellen (Nr. 2b Anhang IV KI-VO-E). Sollte weitergehende Information über das Netz erforderlich sein, so könnte diese ergänzend dargestellt werden – etwa die Architektur als Graph.

[580] Siehe dazu und zum Vorschlag, die Offenlegung der Gewichte an die gestaffelten Regeln zur Offenlegung des Quellcodes zu koppeln, bereits oben, Kapitel 3 C II.

IV. Entscheidungsbäume

Durch einen Entscheidungsbaum werden zumindest diejenigen Merkmale bekannt, die für die Entscheidungsfindung des KNN relevant sind.[581] Es werden jedoch keine präzisen Gewichtungen offengelegt. Die Reproduktion eines funktionsgleichen KNN aus einem Entscheidungsbaum ist daher nicht möglich. Und selbst die Erstellung eines lokalen linearen Modells anhand der Information, die ein Entscheidungsbaum offenbart, erscheint kaum umsetzbar.
Ein Entscheidungsbaum eignet sich mithin hervorragend zur Erfüllung von Transparenzpflichten gegenüber Laien: aus ihm lässt sich intuitiv ablesen, welche Merkmale in eine Entscheidung eingeflossen sind. Nur Informationen über die Gewichtungsverhältnisses lassen sich dem Entscheidungsbaum nicht entnehmen. Gleichzeitig wird nur ein sehr geringer Teil der als Geschäftsgeheimnis geschützten Information offenbar, nämlich allenfalls die Zahl der Eingabe- und Ausgabeneuronen, die sich aus den Merkmalen ablesen lassen.

1. Datenschutzgrundverordnung

Als Information über die involvierte Logik eines KNN im Sinne der Art. 13 Abs. 2 lit. f, Art. 14 Abs. 2 lit. g und Art. 15 Abs. 1 lit. h DSGVO scheint ein Entscheidungsbaum daher durchaus geeignet. Als globale Methode kann er insbesondere auch Einblicke in die abstrakte Funktionsweise des Modells gewähren und mithin auch unproblematisch der Erfüllung der ex-ante-Informationspflichten der Art. 13 und 14 DSGVO dienen. Mit Blick auf beide Parameter, die in der Diskussion um die Reichweite der Informations- und Auskunftsrechte hinsichtlich der involvierten Logik eine Rolle spielen (Verständlichkeit und Geheimnisschutz), wären Entscheidungsbäume daher eine ideale Wahl.

2. Entwurf der KI-Verordnung

Um den im KI-VO-E vorgesehenen Transparenzpflichten nachzukommen, scheint die Methode jedoch weniger gut geeignet.
Als Mittel zur Erklärung der Parameter und ihrer Bedeutung für die Entscheidung gegenüber dem Nutzer in der Gebrauchsanweisung eignen sie sich eher nicht. Denn sie geben keine Information darüber, welches Gewicht einem Merkmal in der Entscheidung des Systems zukommt und dürften es dem Nutzer kaum

[581] Dies müssen jedoch nicht zwingend alle Eingabemerkmale sein.

erlauben „das System angemessen [zu] verstehen und [zu] verwenden“ im Sinne des Art. 13 Abs. 1 KI-VO-E. Mit der Methode wird vielmehr versucht, das Ergebnis eines Blackbox-Modells und die Abhängigkeiten der verschiedenen Merkmale in ein dem Menschen zugängliches Format zu bringen. Aus Sicht des Geheimnisschutzes sind sie zwar den anderen bislang vorgestellten XAI-Methoden vorzuziehen – sie dürften jedoch im Rahmen des Art. 13 KI-VO-E zu ungenau sein.
Dies gilt für die Transparenzpflichten des KI-VO-E gegenüber Behörden erst recht. Denn die technische Dokumentation als Grundlage der Prüfung durch die unterschiedlichen Behörden und Stellen soll Information über die Bedeutung der Parameter enthalten (Nr. 2b Anhang IV KI-VO-E). Zur Überprüfung der Funktionsweise des Modells sind Entscheidungsbäume jedoch zu weit entfernt von der tatsächlichen Entscheidungsfindung und zu ungenau.

V. Kontrafaktische Erklärungen

Aus Kontrafaktischen Erklärungen lässt sich nicht auf die Eigenschaften des zugrundeliegenden Modells schließen und sie sind mithin aus Sicht des Geschäftsgeheimnisschutzes grundsätzlich unbedenklich.[582] Allenfalls ließe sich aus einer Erklärung auf Teile der Daten beziehungsweise Merkmale schließen, mit denen das Modell trainiert wurde. Die Information, dass diese Merkmale genutzt werden, kann als Geschäftsgeheimnis geschützt sein. Sie ist jedoch für das Geschäftsgeheimnis am trainierten KNN unkritisch, denn selbst in Kombination mit weiteren Informationen ist eine Rekonstruktion des KNN anhand ihrer nicht denkbar.

1. Datenschutzgrundverordnung

Kontrafaktische Erklärungen bieten ein sehr einfaches Mittel, Entscheidungen eines KNN für Laien zu erklären.[583] Ihr breiter Einsatz kann helfen, die gesellschaftliche Akzeptanz von ADM zu fördern.[584] Gleichzeitig wird durch sie die „interne Logik“ des Modells nicht preisgegeben, weshalb sie die Voraussetzungen der Transparenzverpflichtungen aus Art. 13 Abs. 2 lit. f, Art. 14 Abs. 2 lit.

[582] So auch *Wachter/Mittelstadt/Russell*, Harvard Journal of Law & Technology 2018, 841 (871).

[583] Siehe dazu *Hacker/Krestel/Grundmann/Naumann*, Artificial Intelligence and Law 2020, 415 (429 f.), m.w.N..

[584] *Wachter/Mittelstadt/Russell*, Harvard Journal of Law & Technology 2018, 841 (843 f.).

g und Art. 15 Abs. 1 lit. h DSGVO grundsätzlich nicht erfüllen dürften.[585] Es könnte jedoch zu überlegen sein, sie im Wege einer teleologischen Auslegung auch hier gelten zu lassen. Denn die Ziele der Erklärung für eine betroffene Person (Entscheidung verstehen, sie angreifen könne sowie verstehen, welches Verhalten die Entscheidung in Zukunft verändern könnte) können mit kontrafaktischen Erklärungen erreicht werden, ohne dass es auf die „interne Logik" im Sinne der DSGVO ankäme.[586]

2. Entwurf der KI-Verordnung

Für die Erfüllung von Transparenzerfordernissen des KI-VO-E sind kontrafaktische Erklärungen ungeeignet. Die Funktionsweise eines Hochrisiko-KI-Systems kann anhand ihrer nicht überprüft werden und die Erklärung ist auch nicht ausreichend, um dem Nutzer „präzise, vollständige, korrekte und eindeutige Informationen" über das KI-System zu geben im Sinne des Art. 13 Abs. 2 KI-VO-E.

VI. Fazit

Mit Ausnahme der kontrafaktischen Erklärungen eignen sich alle untersuchten XAI-Methoden zur Erfüllung von Transparenzpflichten, wenn auch in unterschiedlichem Maße.

Im Rahmen der DSGVO dürften sich insbesondere Diagramme durch LIME und SHAP sowie Entscheidungsbäume eignen, wobei letztere aus Sicht des Geheimnisschutzes die beste Möglichkeit darstellen. Dieses Ergebnis dürfte auch im Einklang mit der höchstrichterlichen Rechtsprechung stehen, wonach eine „Nachrechenbarkeit" der Entscheidung nicht erforderlich ist.[587]

Für die Transparenzpflichten des KI-VO-E könnten insbesondere *Heatmaps*, *Node-Link-Diagramme* und Diagramme durch LIME und SHAP eine gute Möglichkeit darstellen, einen Ausgleich zwischen Transparenz und Geheimnisschutz zu schaffen. Sollte die zur Verfügung gestellte Information im Einzelfall nicht ausreichend sein, bieten sich Ergänzungen durch die Information der dritten Darstellungsstufe an.

---

[585] *Kaminski*, Berkely Tech L. J. 2019, 190 (214); a.A. wohl *Wachter/Mittelstadt/Russell*, Harvard Journal of Law & Technology 2018, 841 (861 ff.).

[586] *Wachter/Mittelstadt/Russell*, Harvard Journal of Law & Technology 2018, 841 (843).

[587] BGH, Urteil v. 28.1.2014, VI ZR 156/13, NJW 2014, 1235, Rn. 25.

D. Urheberrechtlicher Schutz

Die untersuchten XAI-Methoden geben nur in sehr geringem Maße semantische Information des KNN preis, die als Geschäftsgeheimnis geschützt sein könnte. Die Schutzfähigkeit ihrer syntaktischen Information (etwa vor Vervielfältigung) ist daher grundsätzlich von geringer Relevanz uns soll hier nur überblicksartig dargestellt werden.

Die semantische Information des KNN, die als Geschäftsgeheimnis geschützt sein kann, wird in eines *Heatmap* oder der Visualisierung von Merkmalen nicht reproduzierbar offenbar.

Die Visualisierung durch *Heatmaps* und *Feature Visualizations* bleiben so abstrakt, dass eine Erlangung des Geschäftsgeheimnisses an KNN, selbst in Kombination mit den Ergebnissen anderer XAI-Techniken, schwer vorstellbar ist. Und auch in Verbindung mit den Informationen der Darstellungsstufen 3 scheint eine signifikante Erhöhung des Offenbarungsrisikos durch die Offenlegung eines *Heatmap* oder vergleichbarer Visualisierungen nicht wahrscheinlich. Zwar kann der Umstand selbst, dass für eine bestimmte Lösungsfindung eine ML-Technologie eingesetzt wird, ein Geschäftsgeheimnis darstellen. Dieses Geschäftsgeheimnis würde durch die Präsentation eines *Heatmap* zur Erfüllung einer Transparenzpflicht offenbart. Darüber hinaus könnte die kombinierte Offenlegung einer großen Zahl von *Heatmaps* – etwa für jedes Neuron eines KNN – theoretisch Möglichkeiten der Erlangung des Geschäftsgeheimnisses am KNN bieten. Eine Verpflichtung zur Offenlegung einer so großen Menge von *Heatmaps* erscheint aber abwegig – ein Laie könnte aus ihnen keinen Nutzen ziehen.

Auch der urheberrechtliche Schutz von *Heatmaps* und anderen Visualisierungen ist vor dem Hintergrund des geringen Informationsgehalts und des geringen Offenbarungsrisikos kaum relevant. Es käme zwar ein Schutz als Darstellung wissenschaftlicher oder technischer Art gemäß § 2 Abs. 1 Nr. 7 UrhG in Betracht, sofern die Gestaltungshöhe erreicht würde. Allerdings erstreckte sich der Umfang ohnehin nur auf die Form der Darstellung und mithin nicht auf die semantische Information.

Einem *Node-Link-Diagramm* jedoch kann die Architektur des Netzes und mithin sensible Information entnommen werden, die einen Teil des Geschäftsgeheimnisses bildet. Die Schutzmöglichkeiten des Urheberrechts sind, wie auch bei der schematischen Darstellung des KNN als Graph, ungenügend. Denn der

Schutz der konkreten Darstellung hindert nicht das Weiterverbreiten der semantischen Information über den Aufbau des Netzes und mithin womöglich die Offenbarung eines Teils des Geschäftsgeheimnisses.
Die Methoden LIME und SHAP geben allenfalls die verwendeten Merkmale und die Anzahl der Eingabe- und Ausgabeneuronen und mithin nur einen kleinen Teil der als Geschäftsgeheimnis geschützten Information des KNN preis. Eine Erlangung oder gar Offenbarung des Geschäftsgeheimnisses erscheint durch ihre Herausgabe, selbst in Kombination mit der Erlangung weiterer Informationen, nicht realistisch. Die Prüfung weitergehender urheberrechtlicher Schutzmöglichkeiten der Diagramme kann daher dahinstehen.
Gleiches gilt für mit der Hilfe von XAI-Techniken aus KNN generierte Entscheidungsbäume. Aus ihnen lassen sich allenfalls die verwendeten Merkmale und die Anzahl der Neuronen in der Eingabe- und der Ausgabeschicht ablesen. Auch in Kombination mit weiteren Informationen lässt sich damit kein funktionsgleiches KNN nachbauen. Die Frage ihrer Schutzfähigkeit kann mithin für die vorliegende Arbeit ebenfalls dahinstehen.
Kontrafaktische Erklärungen offenbaren mit einzelnen Merkmalen allenfalls Bruchteile der Information eines KNN. Da sie mithin aus Sicht des Geheimnisschutzes unbedenklich sind, stellt sich für sie die Frage einer anderweitigen Schutzmöglichkeit nicht.

E. Fazit

Im Jahr 2018 kam die Gesellschaft für Informatik in einem Gutachten im Auftrag des Sachverständigenrats für Verbraucherfragen hinsichtlich XAI-Methoden und Regulierung zu folgendem Schluss:

> „Die Sichtbarmachung des Vorhersageverhaltens großer neuronaler Netzwerke ist ein Gebiet aktiver Forschung, jedoch nach unserer Einschätzung noch nicht weit genug entwickelt, um regulativ aufgegriffen zu werden."[588]

[588] *Gesellschaft für Informatik*, Technische und rechtliche Betrachtungen algorithmischer Entscheidungsverfahren, S. 55.

Entsprechende Methoden seien daher noch „nicht reif zur verpflichtenden Anwendung“.[589]
Ob diese Einschätzung mittlerweile, über vier Jahre später, anders ausfallen würde, mag dahinstehen. Jedenfalls hat die in diesem Kapitel vorgenommene Analyse gezeigt, welches große Potenzial XAI-Methoden für die Regulierung von Künstlicher Intelligenz haben können.
Denn in ihnen kann sich der Gegensatz zwischen Transparenz und Geheimnisschutz auflösen: sie schaffen eine für den menschlichen Geist verständliche Erklärung der Blackbox, ohne jedoch Geschäftsgeheimnisse zu offenbaren. Gerade interaktive Methoden scheinen daher durchaus geeignet, insbesondere gegenüber Privatpersonen Transparenz der Entscheidungsprozesse von KI-Systemen herzustellen. Die im KI-VO-E vorgesehene digitale Gebrauchsanweisung für KI-Systeme geht daher in eine richtige Richtung und gibt Grund zur Hoffnung, dass entsprechende Techniken regulativ aufgegriffen beziehungsweise Transparenzpflichten von den Gerichten technikoffen und zukunftsgerichtet ausgelegt werden. Ob ein verpflichtender Einsatz von XAI-Methoden zu früh ist, muss hier nicht entschieden werden. Jedenfalls sollte jedoch für Anbieter die Möglichkeit bestehen, diese Methoden zur Erfüllung von Transparenzpflichten zu nutzen. Denn sie werden häufig die einzige Möglichkeit darstellen, diesen Pflichten bei Anwendung von KNN nachzukommen.

[589] *Gesellschaft für Informatik*, Technische und rechtliche Betrachtungen algorithmischer Entscheidungsverfahren, S. 58.

Kapitel 9

Exkurs: Reverse Engineering

Auch außerhalb von Transparenzpflichten stellt sich die Frage, ob die Information eines KNN geheim bleiben kann, wenn die Funktionalität des Netzes einem Nutzer zur Verfügung gestellt wird. Dies kann auf unterschiedliche Weise geschehen, wobei hier der Fokus auf das sog. *Machine-Learning-as-a-service* (*MLaaS*) gelegt wird.[590] Bei *MLaaS* werden ML-Modelle cloud-basiert über eine Schnittstelle (Application Programming Interface, API) zur Verfügung gestellt, über die der Nutzer mit seinen Daten sein eigenes Modell trainieren und anschließend neue eigene Daten klassifizieren kann.[591]
Durch den Zugriff auf das Neuronale Netz wird grundsätzlich die Möglichkeit eröffnet, die als Geschäftsgeheimnis geschützte Information durch eine Rückwärtsanalyse (sog. Reverse Engineering) zu erlangen, die mit der Einführung von GeschGeh-RL und GeschGehG auch in Deutschland ausdrücklich erlaubt worden ist (§ 3 Abs. 1 Nr. 2 GeschGehG). Da die Möglichkeit eines Reverse Engineering bei der Abwägung einer Rolle spielt, ob das immaterialgüterrechtliche Schutzrechtsregime für ML-Modelle angesichts vermehrter Forderungen nach Transparenz ausreichen ist, soll die Thematik hier zumindest kurz umrissen werden. Dazu werden zunächst aktuelle technische Möglichkeiten der Rückwärtsanalyse von KNN besprochen, bevor die Brücke zum Geheimnisschutz geschlagen wird.

---

[590] Eine anderer Anwendungsfall besteht darin, dass KNN als Bestandteil eines IoT-Geräts in den Verkehr gebracht werden; siehe hierzu auch Kapitel 4 B I.

[591] Siehe etwa *Fredrikson/Jha/Ristenpart*, in: Ray/Li/Kruegel, Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security, S. 1322; *Tramèr u. a.*, in: Holz/Savage, Proceedings of the 25th USENIX Security Symposium August 10–12, 2016, S. 601 ff.

A. Technischer Hintergrund

Der Übergang von Methoden der XAI zum Reverse Engineering ist fließend und eine Einteilung kann mit Blick auf die Zielsetzung vorgenommen werden. Als Methoden, die wie die XAI die Erklärung eines Blackbox-Modells zum Ziel haben, können sog. *Nichtinvasive Audits*, *Scraping Audits*, *Sock Puppet Audits* und *Crowdsourced Audits* genannt werden. [592] Diesen Ansätzen ist gemein, dass die erklärungssuchende Person nur als Nutzer auf die Funktion des Modells zugreifen kann, ohne dass ihr eine der in den vorangegangenen Kapiteln beschriebenen Darstellungsformen zur Verfügung stünde. Anhand des Verhaltens des Modells auf unterschiedliche Eingaben, die entweder durch Nutzer (*Nichtinvasive* und *Crowdsourced Audits*) oder durch Skripte bzw. fiktive Benutzer (*Scraping* und *Sock Puppet Audits*) getätigt werden, wird dessen Funktionsweise analysiert. Ziel ist mithin die Erklärung der Funktionsweise beziehungsweise das Auffinden von Fehlern.

Anders verhält es sich bei sogenannten *Model-Extraction*-Techniken, bei denen Parameter und Hyperparameter des Modells herausgefunden werden können, und die häufig eine beinahe vollständige Reproduktion des Modells erlauben.[593] Sie stellen daher eine Art *Reverse Engineering* des Modells dar. Gleichzeitig fällt die Methode nicht mehr unter XAI, wie sie hier verstanden wird. Denn das „kopierte" Modell enthält zwar möglicherweise die (beinahe) identische Information des Originals, diese bleibt jedoch implizit repräsentiert und wird während des *Model-Extraction*-Prozesses nicht explizit.[594]

Die Zielsetzung ist dabei häufig eine andere als bei den genannten Audits: zwar können auch Rückwärtsanalyse und *Model-Extraction* zur Erklärung der Funktionsweise des Modells genutzt werden, die Techniken werden jedoch häufig für

---

[592] *Gesellschaft für Informatik*, Technische und rechtliche Betrachtungen algorithmischer Entscheidungsverfahren, S. 66 ff. Ein Beispiel für ein nichtinvasives bzw. Crowdsourced Audit ist die Kampagne „Open-SCHUFA" von Algorithm Watch und der Open Knowledge Foundation Deutschland.

[593] Siehe eingehend dazu *Oh/Schiele/Fritz*, in: Samek u. a., Explainable AI: Interpreting, Explaining and Visualizing Deep Learning; *Tramèr u. a.*, in: Holz/Savage, Proceedings of the 25th USENIX Security Symposium August 10–12, 2016; *Papernot u. a.*, ASIA CCS 2017, 506; *Ateniese u. a.*, arXiv:1306.4447, 2013.

[594] Auch die Modell-Architektur wird aufgrund der zu lösenden Aufgabe geschätzt, *Papernot u. a.*, ASIA CCS 2017, 506.

„Angriffe“[595] auf *MLaaS*-Modelle verwendet, wie sie mittlerweile von allen großen Internetfirmen angeboten werden.[596] Das Modell stellt für den Nutzer dann eine Blackbox dar, denn er kennt nur die Eingabe und entsprechende Ausgabewerte.[597] Dies ist jedoch ausreichend für eine *Model Extraction* und kann sogar die Anfertigung einer Kopie eines funktionsgleichen Modells ermöglichen. Denn die sich aus Eingabe- und Ausgabewerten ergebende Gleichung kann gelöst und damit die verborgenen Parameter des Modells ergründet werden.[598] Entsprechende “Angriffe” sind dann ohne Überwindung von Zugangsbarrieren oder anderen Sicherheitsvorkehrungen der Anbieter von *MLaaS*-Modellen möglich.[599]
Auch wenn Gegenmaßnahmen getroffen werden können und Gegenstand aktiver Forschung sind[600], so zeigen die Möglichkeiten der „model-extraction-attacks“ doch, dass die Information eines Blackbox-Modells nicht sicher vor Reverse Engineering ist, sofern das Modell den Anwendern zur Klassifizierung eigener Daten zur Verfügung gestellt wird:

> „[...] there is an inherent assumption of secrecy of the ML model in use. We show that this assumption is broken for all ML APIs that we investigate.”[601]

Geschäftsmodelle, die auf das entgeltliche Angebot einer solchen Klassifizierung aufbauen, können so in Gefahr geraten. Denn Angreifer können das Modell kopieren und auf diese Weise Entgelte für weitere Anfragen einsparen (häufig han-

---

[595] In der englischsprachigen Literatur ist meist von „attacks“ die Rede, siehe etwa *Tramèr u. a.*, in: Holz/Savage, Proceedings of the 25th USENIX Security Symposium August 10–12, 2016, S. 601.

[596] *Shokri/Stronati/Song/Shmatikov*, 2017, 1 (2).

[597] Teilweise wird das trainierte Modell auch zum Download angeboten, vgl. *Tramèr u. a.*, in: Holz/Savage, Proceedings of the 25th USENIX Security Symposium August 10–12, 2016, S. 603; dann liegt das KNN dem Nutzer vor (Darstellungsstufe 1) und die Frage der Rückwärtsanalyse stellt sich nicht.

[598] Tramèr et al. sprechen von „equation-solving attacks“, *Tramèr u. a.*, in: Holz/Savage, Proceedings of the 25th USENIX Security Symposium August 10–12, 2016, S. 605.

[599] *Tramèr u. a.*, in: Holz/Savage, Proceedings of the 25th USENIX Security Symposium August 10–12, 2016, S. 610.

[600] Zu Gegenmaßnahmen etwa *Tramèr u. a.*, in: Holz/Savage, Proceedings of the 25th USENIX Security Symposium August 10–12, 2016, S. 614 f.

[601] *Tramèr u. a.*, in: Holz/Savage, Proceedings of the 25th USENIX Security Symposium August 10–12, 2016, S. 604.

delt es sich um sog. *pay-per-query*-Modelle), wodurch der Anbieter möglicherweise seine Trainingskosten nicht mehr amortisieren kann.[602]
Model-Extraction-Techniken ermöglichen also Reproduzierbarkeit, erklären jedoch weder das Modell noch die getroffene Entscheidung und sind daher ungeeignet, Transparenzanforderungen zu erfüllen.

B. Geheimnisschutz

Sowohl Audits als auch *Model-Extraction*-Techniken können auf die eine oder andere Weise Rückschlüsse auf die Funktionsweise des Netzes und gegebenenfalls sogar auf den Maschinencode zulassen.
Dennoch stehen weder der Zugriff des Nutzers über eine Cloud, noch die Übergabe der in einem Gerät integrierten Software dem Geheimnisschutz entgegen, sofern Sie unter der Wahrung eines angemessenen Schutzes erfolgen, also mit Verschlüsselung und gegebenenfalls vertraglicher Absicherung.[603] Denn beide Formen des Zurverfügungstellens des KNN machen weder die Information des KNN „ohne weiteres zugänglich" gemäß § 2 Nr. 1a GeschGehG noch begründen sie für sich genommen Zweifel an der Angemessenheit der Schutzmaßnahmen gemäß § 2 Nr. 1b GeschGehG.
Bei beiden Konstellationen muss jedoch beachtet werden, dass die Rückwärtsanalyse, das sog. *Reverse Engineering*, nach § 3 Abs. 1 Nr. 2 GeschGehG explizit erlaubt ist für Gegenstände, die öffentlich verfügbar gemacht wurden (lit. a) oder sich im rechtmäßigen Besitz eines Beobachtenden befinden, der keiner Pflicht zur Beschränkung der Erlangung des Geschäftsgeheimnisses unterliegt (lit. b). Auch wenn die Information des KNN somit ein Geschäftsgeheimnis darstellt nach § 2 Nr. 1 GeschGehG, so kann ihre Erlangung durch eine nach § 3 Abs. 1 Nr. 2 GeschGehG erlaubte Handlung gestattet sein.[604] Kann daher auf ein KNN über eine Cloud von jedermann zugegriffen werden, auch gegen Be-

[602] *Tramèr u. a.*, in: Holz/Savage, Proceedings of the 25th USENIX Security Symposium August 10–12, 2016, S. 606.

[603] Siehe zu Sicherheitsproblemen, wenn Web-Anwendungen über eine Programmierschnittstelle bereitgestellt werden *Deutsch/Eggendorfer*, in: Taeger/Pohle, Computerrechts-Handbuch : Computertechnologie in der Rechts- und Wirtschaftspraxis, 50.1 Rn. 80.

[604] Siehe dazu nur *Hauck/Cevc*, ZGE 2019, 135 (166 f.), vorbehaltlich des Nachahmungsschutzes in § 4 Nr. 3 UWG.

zahlung, so ist von seiner öffentlichen Verfügbarkeit i. S. d. § 3 Abs. 1 Nr. 2a GeschGehG auszugehen.[605] Auch ein Registrierungserfordernis dürfte dem wohl nicht entgegenstehen. Es kommt darauf an, ob das KNN für die Allgemeinheit zugänglich ist, selbst wenn der Zugang mit großen Mühen oder hohen Kosten verbunden ist.[606] Ist das KNN danach öffentlich verfügbar, kann auch keine individualvertragliche Untersagung einer Rückwärtsanalyse erfolgen.[607]
Dies ergibt sich bereits im Umkehrschluss aus § 3 Abs. 1 Nr. 2b GeschGehG.[608] Nach diesem kann die Rückwärtsanalyse vertraglich ausgeschlossen werden, sofern sich das KNN im rechtmäßigen Besitz eines Beobachtenden befindet und, so der Umkehrschluss aus lit. a, nicht öffentlich verfügbar gemacht wurde. Vorstellbar ist dies etwa bei der Übergabe im Rahmen eines Kaufvertrages über ein IoT-Gerät, oder aber im Rahmen von Software-Lizenzen.[609]
Bei solchen Massengeschäften drängt sich jedoch die Frage der Wirksamkeit einer entsprechenden Beschränkung in AGB auf. Da die Gestattung des Reverse Engineering bei jedem rechtmäßigem Besitz beträchtlich ist und § 3 Abs. 1 Nr. 2b GeschGehG eine Beschränkung ausdrücklich vorsieht, verstieße eine solche Klausel wohl nicht gegen § 307 Abs. 2 Nr. 1 BGB. Anders jedoch in dem Fall, dass die Rückwärtsanalyse in anderen Vorschriften explizit gestattet ist. Die Dekompilierung eines Computerprogramms etwa kann nicht wirksam über AGB unter Berufung auf § 3 Abs. 1 Nr. 2b GeschGehG ausgeschlossen werden, da sie in § 69d Abs. 2, 3, § 69e UrhG erlaubt ist.[610]
Mit § 3 Abs. 1 Nr. 2 GeschGehG wird jedoch zunächst nur die Erlangung des Geschäftsgeheimnisses erlaubt. Die Vorschrift erlaubt weder dessen Nutzung oder Offenlegung, noch sagt sie etwas über das Schicksal des Geschäftsgeheimnisses aus. Das Reverse Engineering selbst hat damit zunächst keinen Einfluss

---

[605] So ausdrücklich für „digitale Gegenstände wie Computerprogramme oder Daten" *Ohly*, in: Harte-Bavendamm/Ohly/Kalbfus, GeschGehG, § 3 Rn. 24.

[606] *Ohly*, in: Harte-Bavendamm/Ohly/Kalbfus, GeschGehG, § 3 Rn. 24.

[607] So im Ergebnis auch *Ohly*, in: Harte-Bavendamm/Ohly/Kalbfus, GeschGehG, § 3 Rn. 25; anders möglicherweise Kuss/Sassenberg, die jedoch nicht nach den Fällen des § 3 Abs. 1 Nr. 2 a und b GeschGehG differenzieren: *Kuß/Sassenberg*, in: Sassenberg/Faber, Rechtshandbuch Industrie 4.0 und Internet of Things: Praxisfragen und Perspektiven der digitalen Zukunft, S. 449.

[608] So auch *Ohly*, in: Harte-Bavendamm/Ohly/Kalbfus, GeschGehG, § 3 Rn. 25.

[609] *Surblyté*, in: Ullrich/Hilty/Lamping/Drexl, TRIPS plus 20: From Trade Rules to Market Principles, S. 743.

[610] Siehe insgesamt zur Frage der wirksamen Beschränkung der Rückwärtsanalyse im Rahmen von AGB *Ohly*, in: Harte-Bavendamm/Ohly/Kalbfus, GeschGehG, § 3 Rn. 27.

auf den Bestand des Geschäftsgeheimnisses. Die Information wird durch die bloße Erlangung nicht offenkundig, solange sie nicht offengelegt wird.[611]
Ob eine solche Offenlegung oder auch Nutzung nach erlaubter Rückwärtsanalyse gestattet ist, richtet sich nach § 4 Abs. 2 Nr. 2 GeschGehG: es kommt mithin darauf an, ob diesbezüglich einer Verpflichtung zuwidergehandelt wird. Ist dies der Fall, so stehen dem Geheimnisinhaber die in den §§ 6 ff. GeschGehG niedergelegten Ansprüche u. a. auf Beseitigung und Unterlassung, Vernichtung und Schadenersatz zu. Bei Vorliegen weiterer Voraussetzungen kann die entsprechende Handlung sogar nach § 23 GeschGehG strafbar sein.
Die im GeschGehG kodifizierte Erlaubnis des Reverse Engineering verdeutlicht, wie wichtig die differenzierte Betrachtung der verschiedenen Informationen eines KNN sowie ihrer Darstellungsformen ist. Selbst wenn das KNN „öffentlich verfügbar gemacht wurde" gemäß § 3 Abs. 1 Nr. 2a GeschGehG, bedeutet das nicht, dass der Maschinencode öffentlich verfügbar ist. Wird ein Produkt oder Gegenstand öffentlich verfügbar gemacht, so kann die darin enthaltene Information daher dennoch geheim bleiben. Es besteht lediglich ein größeres Risiko, dass er mittels Techniken der Rückwärtsanalyse erlangt und möglicherweise, rechtmäßig oder rechtswidrig, offengelegt wird.
Aus urheberrechtlicher Sicht stellen entsprechende Techniken dann keine Verletzung des Rechts an der Software dar, wenn sie sich in den Grenzen der Schrankenbestimmung des § 69d Abs. 3 UrhG halten. Denn diese umfasst auch den Zweck, das Know-how des Herstellers zu erlangen.[612]

[611] *Hauck/Cevc*, ZGE 2019, 135 (167); *Ohly*, in: Harte-Bavendamm/Ohly/Kalbfus, GeschGehG, § 3 Nr. 29.

[612] *Triebe*, WRP 2018, 795 (798 f.); siehe grundlegend zur urheberrechtlichen Bewertung der Ermittlung von Ideen und Grundsätzen des Programms EuGH, Urteil v. 2.5.2012, C-406/10 – *SAS Institute*.

Teil 4

Die Symbiose von Erklärbarkeit und Geheimnisschutz: Abgestufte Transparenz

Die Information eines KNN ist nach geltendem Immaterialgüterrecht nur eingeschränkt geschützt. Dies gilt sowohl für die semantische als auch für die syntaktische Information. Denn KNN werden regelmäßig nicht dem Schutz durch das Patentrecht zugänglich sein und der Geschäftsgeheimnisschutz ist als rein faktischer Schutz risikobehaftet. Sie unterfallen außerdem in ihren wesentlichen Teilen nicht dem Urheberrecht.

Es drängt sich daher die Frage auf, ob die Schaffung eines neuen Schutzrechts beziehungsweise eine Veränderung bestehender Schutzrechte erforderlich ist.

Die Zusammenführung der Ergebnisse der vorangegangenen Analyse wird jedoch zeigen, dass der Schutz von KNN auch im bestehenden immaterialgüterrechtlichen System ausreichend gewährleistet werden kann.[613] Und dies selbst angesichts zahlreicher – und voraussichtlich zunehmender – Transparenzverpflichtungen. Denn durch den Rückgriff auf ein differenziertes Darstellungssystem der Information von KNN kann die Offenbarung von Geschäftsgeheimnissen vermieden werden. Die Wahl der passenden Darstellungsform kann dann abgestuft nach Empfängertyp und Sinn und Zweck der Transparenzpflicht erfolgen. Durch ein solches *System abgestufter Transparenz* wird es auf eine Abwägung von Offenbarungsinteressen und Geschäftsgeheimnisschutz dann in den meisten Fällen nicht einmal ankommen.

Die beispielhaft untersuchten Transparenzpflichten sehen verschiedene Empfänger von Information vor, die grob in öffentliche[614] und private Empfänger unterteilt werden können. Bei den Privaten lässt sich dann noch unterscheiden zwischen solchen, die in ihrer Eigenschaft als Privatperson von KI betroffen sind (Verbraucher im Sinne von § 13 BGB) und solchen, welche entweder gewerblich von KI betroffen sind oder selbst KI-Produkte gewerblich verwenden. Diese drei Empfängergruppen dürften sich auch in anderen bestehenden und zukünftigen Transparenzpflichten unterscheiden lassen.

---

[613] Die Einschätzung, dass ein neues Schutzrecht nicht erforderlich ist, deckt sich mit den überwiegenden Stimmen in der Forschung. Dort wird die Frage jedoch fast ausschließlich unabhängig von der Transparenzfrage betrachtet: *Hauck/Cevc*, ZGE 2019, 135 (168); *Hartmann/Prinz*, in: Taeger, Rechtsfragen digitaler Transformationen - Gestaltung digitaler Veränderungsprozesse durch Recht, S. 787 f., allerdings aufgrund der Annahme, dass ausreichender urheberrechtlicher Schutz gegeben; unentschlossen: *Wissenschaftliche Dienste des Deutschen Bundestags*, Künstliche Intelligenz und Machine Learning, S. 23; *Tochtermann*, in: Kaulartz/Braegelmann, Rechtshandbuch Artificial Intelligence und Machine Learning, S. 326 ff.

[614] Im Folgenden als „öffentliche Stellen“ bezeichnet. Damit sind hier Behörden und auch Dritte gemeint, die mit der Überprüfung von Information durch Behörden betraut werden, unabhängig vom Vorliegen einer Beleihung.

Kapitel 10

Transparenz gegenüber öffentlichen Informationsempfängern

Die Transparenzpflichten gegenüber öffentlichen Stellen sind naturgemäß die umfangreichsten, wie sich am jüngsten und ersten Beispiel umfassender KI-Regulierung, dem KI-VO-E, zeigt. Denn öffentliche Stellen müssen im Zweifel die Konformität von KI-Anwendungen mit wichtigen Schutznormen überprüfen können.

Ihnen gegenüber wird daher häufig das Geschäftsgeheimnis an einem Künstlichen Neuronalen Netz zu offenbaren sein, das heißt die Gewichte und die Systemarchitektur, gegebenenfalls auch Quell- oder sogar Maschinencode.

Dies ist jedoch in Anbetracht der Verschwiegenheitsverpflichtungen dieser Empfängergruppe verhältnismäßig unproblematisch.[615] Je weiter sich die Offenbarung sensibler, geheimer Information allerdings von einer zentralen, mit Audits betrauten Behörde entfernt – durch Zugriffsrechte weiterer Behörden, Beauftragung Dritter mit der Überprüfung oder sogar die Einräumung weiterer Subdelegationsrechte an diese Dritte – desto kritischer wird dies für den Geheimnisinhaber. Hier sollten daher gestaffelte Zugriffsrechte beziehungsweise Offenbarungspflichten vorgesehen werden. Quellcode und insbesondere Gewichte eines KNN sollten konzentriert von einer öffentlichen Stelle überprüft werden und für weitere öffentliche Stellen sollte die Möglichkeit eingeräumt werden, eine entsprechende Überprüfung zu beantragen. Dies entspricht der im KI-VO-E hinsichtlich des Quellcodes getroffenen Regelung, die im Bereich von KNN insbesondere auch für die Gewichte gelten sollte.

Angesichts der Herausforderungen, vor die sogenannte Code-Audits im Bereich des Maschinellen Lernens gestellt sind, sollte darüber hinaus erwogen werden, ob eine Herausgabe von Quell- oder Maschinencode überhaupt das geeignete Mittel

[615] So auch *Martini*, Blackbox Algorithmus – Grundfragen einer Regulierung Künstlicher Intelligenz, S. 254.

zur Überprüfung der Funktionsweise eines KNN darstellt.[616] So werden die im vorangegangenen Kapitel beschriebenen Testungen als „zielführendste" Methode angesehen, um Diskriminierungen beim Einsatz von ADM gegenüber Verbrauchern aufzudecken, die obendrein ohne Offenlegung von Geschäftsgeheimnissen möglich ist.[617]

[616] Siehe dazu *Gesellschaft für Informatik*, Technische und rechtliche Betrachtungen algorithmischer Entscheidungsverfahren, S. 141.

[617] *Gesellschaft für Informatik*, Technische und rechtliche Betrachtungen algorithmischer Entscheidungsverfahren, S. 58.

Kapitel 11

Transparenz gegenüber privaten Informationsempfängern

Die aus Sicht des Geheimnisträgers kritischste Gruppe von Informationsempfängern dürften die Unternehmer sein. Denn ihnen gegenüber greifen einerseits manche der gegenüber Verbrauchern geltenden (faktischen) Restriktionen nicht immer, wie das Erfordernis der Verständlichkeit. Andererseits werden unter ihnen auch häufig Wettbewerber des von der Transparenzpflicht betroffenen Geheimnisträgers sein, die ein großes Interesse an den Funktionalitäten des KNN haben können.

Ausgerechnet gegenüber dieser kritischen Empfängergruppe sind die Transparenzverpflichtungen des KI-VO-E unklar ausgestaltet. Während eine Offenbarung von Quell- oder Maschinencode ihr gegenüber nicht vorgesehen ist, ist die Verpflichtung zur Offenlegung der Gewichte eines KNN nach dem aktuellen Wortlaut zumindest nicht ausgeschlossen. Da die Gewichte jedoch das Herzstück der Information eines KNN ausmachen, wäre eine eindeutigere Regelung wünschenswert. Jedenfalls im Rahmen einer Interessenabwägung muss hier dem Geheimhaltungsinteresse gegenüber dem Offenbarungsinteresse Vorrang eingeräumt werden. Dies gilt umso mehr, als das Ziel, das mit den Transparenzpflichten gegenüber Nutzern verfolgt wird, mit aus Sicht des Geheimnisschutzes wesentlich unkritischerer Information erfüllt werden kann. Nutzer sollen „präzise, vollständige, korrekte und eindeutige Informationen in einer [...] relevanten, barrierefrei zugänglichen und verständlichen Form“ erhalten (Art. 13 Abs. 2 KI-VO-E) und „das System angemessen verstehen und verwenden können“ (13 Abs. 1 KI-VO-E). Anders als für öffentliche Stellen zielt die Regelung mithin nicht darauf ab, die Überprüfung der Funktionalität des Systems zu ermöglichen. Das Verständnis des Systems und die darauf aufbauende Verwendung kann aber durch die Darstellungsformen der dritten und vierten Stufe, auch in Kombination, „angemessen“ erreicht werden im Sinne des Art. 13 Abs. 1 KI-VO-E. So könnte etwa der Aufbau des Netzes durch die schematische Darstellung als Graph oder in natürlicher Sprache erklärt werden, ohne dass hierbei die Ge-

wichte offenbart werden müssten. Dem eigentlichen Kernanliegen des Art. 13 KI-VO-E, Verständnis und Verwendung des Systems, kann etwa durch ein SHAP-Diagramm vollumfänglich nachgekommen werden, insbesondere durch eine interaktive Variante. Diese Darstellungsmethoden stellen einen im Sinne des Art. 13 Abs. 1 KI-VO-E durchaus „angemessenen" Weg dar, den widerstreitenden Interessen von Geheimnisinhaber und Nutzer in gleichem Maße Rechnung zu tragen.

Auch aus der DSGVO können sich theoretisch Transparenzpflichten von Unternehmern gegenüber potenziellen Konkurrenten ergeben. Die „betroffene Person" (vgl. Artikel 4 Nummer 1 DSGVO), die ein Recht auf Information über die Erhebung und auf Auskunft über die Verarbeitung ihrer personenbezogenen Daten hat, kann zwar nur eine natürliche Person, jedoch auch eine beruflich handelnde natürliche Person sein.[618] Die vorherrschende Gruppe von nach Artikel 13 bis 15 DSGVO Informations- und Auskunftsberechtigten dürften jedoch die Verbraucher sein.

Der Umfang der sich aus der DSGVO ergebenden Informations- und Auskunftsrechte ist, wie gezeigt, umstritten. Geheimnisschutz und Datenschutz stehen sich hier konfliktgeladen gegenüber und das Gesetz selbst sieht keine Lösung, nicht einmal einen methodischen Ansatz zu einer Lösung dieses Konflikts vor.

Zumindest für Informationen über Datenverarbeitung durch Künstliche Neuronale Netze ließe sich dieser Konflikt mit wenigen Reibungen auflösen. Geheimnisschutz und Verständlichkeitsgebot (Artikel 12 DSGVO) spannen dabei den Bereich auf, in dem sich diese Lösung finden lässt. Denn das Verständlichkeitsgebot dient hier (ungewollt) auch dem Geheimnisschutz. Zwar kann unter Berufung auf die Verständlichkeit dem Betroffenen nicht jegliche Offenbarung auch einer komplexeren „involvierten Logik" abgesprochen werden. Dennoch kommt dem Geheimnisschutz hier gewissermaßen die Eigenheit der Information Künstlicher Neuronaler Netze zu Gute: die Informationsverarbeitung des Netzes wird in den traditionellen Darstellungsformen (hier Stufen 1 bis 3) nicht explizit, selbst Experten sprechen von einer Blackbox. Die verständliche Darstellung der Informationsverarbeitung eines KNN ist auf diese Art nicht möglich – Erklärbarkeit und Geheimnisschutz bilden eine Symbiose. Erst die Darstellungsmöglichkeiten der vierten Stufe bieten die Möglichkeit, die potentielle Information über die Informationsverarbeitung im Netz zu faktischer Information werden zu lassen.

---

[618] Siehe nur *Ernst*, in: Paal/Pauly DS-GVO, Art. 4 Rn. 4.

Sofern daher mit einer äußerst extensiven Auslegung der Artikel 13 bis 15 DSGVO eine Offenbarung des Quellcodes – oder sogar der Gewichte – gefordert wird, ist dies für KNN aus zweierlei Gründen abzulehnen. Zum einen betreffen zumindest die Gewichte den Kern der als Geschäftsgeheimnis schutzfähigen Information eines KNN. Mit der nach der DSGVO geforderten Wahrung von Geschäftsgeheimnissen wäre eine Offenlegung daher nicht zu vereinbaren. Denn auch im Rahmen einer Interessenabwägung ist nicht ersichtlich, welche überwiegenden Interessen der betroffenen Person einen derartigen Eingriff in die Rechte des Geheimnisträgers rechtfertigen könnten. Hier kommt dem Geheimnisschutz nämlich das Verständlichkeitsgebot zur Hilfe: angesichts der bestehenden Möglichkeiten, die Information eines KNN auf verständliche Weise darzustellen, ist nicht ersichtlich, warum ein Betroffener Zugang zu sensibler Information bekommen sollte, die für ihn in den meisten Fällen absolut unverständlich sein dürfte.

Auch zur Erfüllung der Informations- und Auskunftspflichten aus der DSGVO bietet sich daher eine Kombination der Darstellungsmethoden aus der dritten und vierten Stufe an. Mehr noch als gegenüber den Nutzern im Sinne des KI-VO-E dürfte der Fokus hier auf die vierte Stufe zu legen sein. Besonders geeignet zur Erklärung der involvierten Logik einer Entscheidung sind *Heatmaps*, LIME- und SHAP-Diagramme und Entscheidungsbäume. Diese Methoden erlauben es der betroffenen Person, die wesentlichen Parameter einer Entscheidung abzulesen. Mit Ausnahme der Entscheidungsbäume ermöglichen sie es sogar zu erkennen, welches Gewicht dem jeweiligen Parameter in der Entscheidung zukam, ohne jedoch die dahinterliegenden Gewichte des KNN zu offenbaren.[619] Auch hier könnten interaktive Methoden helfen, die Hintergründe einer automatisierten Entscheidung zu verstehen. Dies könnte auch und gerade in Kombination mit kontrafaktischen Erklärungen sinnvoll sein. Zwar ist durchaus fraglich ist, ob diese für sich genommen die Informations- und Auskunftsrechte der betroffenen Person hinsichtlich der involvierten Logik befriedigen können. Stünde jedoch etwa ein interaktives SHAP-Diagramm zur Verfügung, in dem die betroffene Person die Werte der relevanten Parameter selbst verändern und ihren Einfluss auf die Entscheidung nachvollziehen kann, so könnte das Potenzial kontrafaktischer Erklärungen ausgeschöpft werden.

[619] Auch Hacker schlägt daher für einen Ausgleich zwischen Transparenz und Geheimhaltung eine „unpräzise Offenlegung der numerischen Gewichte bei gleichzeitiger voller Nennung der entsprechenden features“ vor: *Hacker*, NJW 2020, 2142 (2144).

Kapitel 12

System abgestufter Transparenz

Die vorangegangenen Kapitel haben gezeigt, dass die detaillierte Analyse der in einem KNN enthaltenen Informationen und ihrer Darstellungsformen den Entwurf eines *Systems abgestufter Transparenz* ermöglicht, das sich in das bestehende Regulierungssystem integriert.[620]

Dieses System dürfte grundsätzlich auch auf andere bestehende und zukünftige Regelungswerke zur Transparenz von KI-Systemen übertragbar sein. Denn Transparenzpflichten werden üblicherweise im Sinne einer Abstraktheit und Technikoffenheit des Rechts verhältnismäßig vage ausgestaltet sein und somit eine Auslegung nach den dargestellten Prinzipien abgestufter Transparenz erlauben. Auch eine Verhältnismäßigkeitsprüfung, wie sie etwa Art. 3 Abs. 4 KI-Haftungs-RL-E vorsieht, kann durch das System abgestufter Transparenz erleichtert werden.

Dies dürfte sogar in den Fällen gelten, in denen Informationen veröffentlicht werden müssen und daher nicht einmal die Möglichkeit von Geheimhaltung oder vertraglicher Einschränkung besteht. Teilweise wird zwar argumentiert, dass durch entsprechend weitgehende Normen die Abwägung zwischen den widerstreitenden Interessen bereits zugunsten der Transparenz entschieden worden sei, da Geschäftsgeheimnisse mit Veröffentlichung der Information bereits tatbestandlich entfielen. So sieht Alexander mit Blick auf die P2B-Verordnung[621] den „dogmatisch richtige[n] Standort im Recht der Geschäftsgeheim-

[620] Die abgestufte Transparenz integriert damit auch die von Hacker geforderte „qualifizierte[...] Transparenz", wonach geheime Information nur gegenüber zur Verschwiegenheit verpflichteten Empfängern offengelegt werden soll. Vgl. *Hacker*, NJW 2020, 2142 (2144).

[621] Verordnung (EU) 2019/1150 des Europäischen Parlaments und des Rates vom 20. Juni 2019 zur Förderung von Fairness und Transparenz für gewerbliche Nutzer von Online- Vermittlungsdiensten.

nisse für die Lösung des Konfliktbereichs zwischen Informationspflichten und dem Schutz von Geschäftsgeheimnissen [in] § 2 Nr. 1 Gesch-GehG."[622]
Dem kann jedoch nach der hier vorgenommenen differenzierten Analyse der geheimen Information und der Geschäftsgeheimnisse an einem KNN nicht vollumfänglich zugestimmt werden. Natürlich wird durch Veröffentlichung geheimer Information diese „allgemein zugänglich" im Sinne des § 2 Nr. 1a Gesch-GehG und ein Geschäftsgeheimnis liegt schon tatbestandlich nicht vor. Allerdings bedarf es einer differenzierten Betrachtung, welche Information eines KI-Systems zur Verfügung gestellt werden muss, ob diese überhaupt geheim ist und ob mit ihr auch weitere Teile des Geschäftsgeheimnisses am ganzen System gefährdet sind.[623] Müsste der Anbieter eines Ranking-Algorithmus etwa lediglich die verwendeten Parameter und eine ungefähre Gewichtung öffentlich preisgeben, so dürfte diese Information nur einen verhältnismäßig kleinen Teil seines Geschäftsgeheimnisses an dem (im Falle von Rankings wohl üblicherweise) verwendeten KNN darstellen.[624] Auch so weitgehende Transparenzpflichten wie Artikel 5 Abs. 1 und 2 P2B-VO fordern zudem nur die Veröffentlichung der „relative[n] Gewichtung" von Hauptparametern gegenüber anderen Parametern, womit nicht die Offenlegung von allen Gewichten eines KNN gemeint sein kann, zumal die Information (zumindest in Absatz 2) durch „klar und verständlich formulierte Erläuterungen" bereitgestellt werden soll. Ob zumindest die wichtigen Ranking-Parameter nicht „den Personen in den Kreisen, die üblicherweise mit dieser Art von Informationen umgehen, allgemein bekannt" sind und mithin unabhängig von ihrer Veröffentlichung schon tatbestandlich als Geschäftsgeheimnis ausscheiden, dürfte ebenfalls zu überlegen sein.
Die nach Darstellungsformen und Empfängerhorizont abgestufte Transparenz bietet eine Möglichkeit, die diffizile und einzelfallbezogene Abwägung zwischen Geheimhaltungs- und Offenlegungsinteressen in vielen Fällen zu vermeiden, indem die Frage nach dem Umfang der Offenlegung bereits tatbestandlich – auf der Ebene des Offenlegungsobjekts – entschieden wird. Auf diese Weise wird den Interessen der unterschiedlichen Akteure angemessen Rechnung getragen.

---

[622] *Alexander*, MMR 2021, 690 (695).

[623] Auch Alexander fordert daher die Prüfung, in welchem Umfang Geschäftsgeheimnisse von einer Informationspflicht betroffen sind. *Alexander*, MMR 2021, 690 (695).

[624] Anders Alexander, für den Transparenzpflichten für Ranking-Systeme in einem „Exklusivitätsverhältnis" zum Schutz von Geschäftsgeheimnissen stehen, *Alexander*, MMR 2021, 690 (694).

Die „holzschnittartige Lösung, wonach Aufklärungsbemühungen bei fadt [sic.] jeder Betroffenheit von Betriebs- und Geschäftsgeheimnissen weitgehend ins Leere laufen“[625] wird durch ein differenziertes und biegsames Repertoire an Lösungsmöglichkeiten ersetzt. Angesichts dieses Repertoires und der zu erwartenden Fortschritte im Bereich der XAI kann die Notwendigkeit eines neuen Schutzrechts für KNN oder ihre Bestandteile auch angesichts umfassender Transparenzpflichten nicht festgestellt werden.

[625] *Wischmeyer*, AöR 2018, 1 (64).

Ergebnisse

Künstliche Neuronale Netze werden häufig als Blackbox bezeichnet. Ihre Funktionsweise wird im ersten Teil dieser Arbeit jedoch detailliert beschrieben, was eine anschließende Annäherung an ihre semantische Information ermöglicht. Dafür werden im zweiten Teil verschiedene Informationsbegriffe herangezogen, wobei insbesondere die Unterscheidung von impliziter und expliziter, von semantischer und syntaktischer sowie von Information im Sinne des Geschäftsgeheimnisgesetzes von Interesse ist. Die Information des KNN liegt im Spannungsverhältnis zwischen Geheimnisschutz und Transparenz, das sich anhand der Transparenzpflichten in der Datenschutzgrundverordnung und dem Entwurf der KI-Verordnung nachvollziehen lässt. In beiden Regelungswerken bleiben jedoch sowohl der konkrete Umfang der Transparenzpflichten als auch eine daran anknüpfende Lösung des Interessenkonflikts zwischen Geheimnisinhaber und Informationsempfänger offen.

Der Umfang der geforderten Transparenz kann für Künstliche Neuronale Netze im dritten Teil jedoch detailliert herausgearbeitet werden. Dazu muss zunächst die semantische Information des Netzes bestimmt werden. Sie besteht in den Regeln, nach denen das Netz Information verarbeitet. Immaterialgüterrechtlicher Schutz der semantischen Information kann im Wesentlichen über den Schutz als Geschäftsgeheimnis erfolgen.

Es bestehen dann unterschiedliche Möglichkeiten, diese semantische Information eines KNN als syntaktische Information darzustellen. Diese Möglichkeiten lassen sich nach ihrer Abstraktheit als vier Stufen beschreiben, wobei die erste Stufe die konkreteste, die vierte Stufe die abstrakteste Darstellung enthält. Der Maschinencode steht auf der ersten Stufe, da er die gesamte semantische Information des KNN enthält. Es folgen Quellcode und Gewichtsdatei auf der zweiten und Graph, natürliche Sprache und mathematische Formel auf der dritten Stufe. Auf der vierten Stufe steht die Art von Information, wie sie durch Techniken der *Explainable Artificial Intelligence* generiert wird.

Die Untersuchung der syntaktischen Information jeder Stufe liefert dann unterschiedliche Ergebnisse. Während manche Darstellungsmöglichkeiten die (na-

hezu) vollständige semantische Information des KNN enthalten (Maschinencode, Quellcode mit Gewichten, Gewichte) und ihre Offenlegung daher das Risiko eines Geheimnisverlusts birgt, ist für andere eine differenzierte Analyse des Informationsgehalts erforderlich (Quellcode allein, Darstellungsmöglichkeiten der dritten Stufe). Für die Darstellungsmöglichkeiten der vierten Stufe ist charakteristisch, dass sie einerseits die Funktionsweise eines KNN explizit darstellen können, andererseits die semantische Information des Netzes nicht reproduzierbar offenbaren.

Die Möglichkeiten urheberrechtlichen Schutzes sind für die Darstellungen aller Stufen nur begrenzt.

Die Zusammenführung der Untersuchung im vierten Teil zeigt, dass durch ein *System abgestufter Transparenz* den Interessen von Informationsempfängern und Geheimnisinhabern angemessen Rechnung getragen und das Spannungsverhältnis zwischen Transparenz und Geheimnisschutz aufgelöst werden kann.

Anhang: Auszüge aus analysierten Regelungswerken

Gesetz zum Schutz von Geschäftsgeheimnissen[626]

Abschnitt 1 Allgemeines

§ 1 Anwendungsbereich

(1) Dieses Gesetz dient dem Schutz von Geschäftsgeheimnissen vor unerlaubter Erlangung, Nutzung und Offenlegung.
(2) Öffentlich-rechtliche Vorschriften zur Geheimhaltung, Erlangung, Nutzung oder Offenlegung von Geschäftsgeheimnissen gehen vor.
(3) Es bleiben unberührt:
    1. der berufs- und strafrechtliche Schutz von Geschäftsgeheimnissen, deren unbefugte Offenbarung von § 203 des Strafgesetzbuches erfasst wird,
    2. die Ausübung des Rechts der freien Meinungsäußerung und der Informationsfreiheit nach der Charta der Grundrechte der Europäischen Union (ABl. C 202 vom 7.6.2016, S. 389), einschließlich der Achtung der Freiheit und der Pluralität der Medien,
    3. die Autonomie der Sozialpartner und ihr Recht, Kollektivverträge nach den bestehenden europäischen und nationalen Vorschriften abzuschließen,
    4. die Rechte und Pflichten aus dem Arbeitsverhältnis und die Rechte der Arbeitnehmervertretungen.

§ 2 Begriffsbestimmungen

Im Sinne dieses Gesetzes ist
    1. Geschäftsgeheimnis
       eine Information
        a) die weder insgesamt noch in der genauen Anordnung und Zusammensetzung ihrer Bestandteile den Personen in den Kreisen, die üb-

[626] Zum 25.07.2023 aktuellste verfügbare Fassung der Gesamtausgabe

licherweise mit dieser Art von Informationen umgehen, allgemein bekannt oder ohne Weiteres zugänglich ist und daher von wirtschaftlichem Wert ist und
   b) die Gegenstand von den Umständen nach angemessenen Geheimhaltungsmaßnahmen durch ihren rechtmäßigen Inhaber ist und
   c) bei der ein berechtigtes Interesse an der Geheimhaltung besteht;
2. Inhaber eines Geschäftsgeheimnisses
   jede natürliche oder juristische Person, die die rechtmäßige Kontrolle über ein Geschäftsgeheimnis hat;
3. Rechtsverletzer
   jede natürliche oder juristische Person, die entgegen § 4 ein Geschäftsgeheimnis rechtswidrig erlangt, nutzt oder offenlegt; Rechtsverletzer ist nicht, wer sich auf eine Ausnahme nach § 5 berufen kann;
4. rechtsverletzendes Produkt
   ein Produkt, dessen Konzeption, Merkmale, Funktionsweise, Herstellungsprozess oder Marketing in erheblichem Umfang auf einem rechtswidrig erlangten, genutzten oder offengelegten Geschäftsgeheimnis beruht.

§ 3 Erlaubte Handlungen

(1) Ein Geschäftsgeheimnis darf insbesondere erlangt werden durch
1. eine eigenständige Entdeckung oder Schöpfung;
2. ein Beobachten, Untersuchen, Rückbauen oder Testen eines Produkts oder Gegenstands, das oder der
   a) öffentlich verfügbar gemacht wurde oder
   b) sich im rechtmäßigen Besitz des Beobachtenden, Untersuchenden, Rückbauenden oder Testenden befindet und dieser keiner Pflicht zur Beschränkung der Erlangung des Geschäftsgeheimnisses unterliegt;
3. ein Ausüben von Informations- und Anhörungsrechten der Arbeitnehmer oder Mitwirkungs- und Mitbestimmungsrechte der Arbeitnehmervertretung.

(2) Ein Geschäftsgeheimnis darf erlangt, genutzt oder offengelegt werden, wenn dies durch Gesetz, aufgrund eines Gesetzes oder durch Rechtsgeschäft gestattet ist.

§ 4 Handlungsverbote

(1) Ein Geschäftsgeheimnis darf nicht erlangt werden durch
1. unbefugten Zugang zu, unbefugte Aneignung oder unbefugtes Kopieren von Dokumenten, Gegenständen, Materialien, Stoffen oder elektronischen Dateien, die der rechtmäßigen Kontrolle des Inhabers des Geschäftsgeheimnisses unterliegen und die das Geschäftsgeheimnis enthalten oder aus denen sich das Geschäftsgeheimnis ableiten lässt, oder
2. jedes sonstige Verhalten, das unter den jeweiligen Umständen nicht dem Grundsatz von Treu und Glauben unter Berücksichtigung der anständigen Marktgepflogenheit entspricht.

(2) Ein Geschäftsgeheimnis darf nicht nutzen oder offenlegen, wer
1. das Geschäftsgeheimnis durch eine eigene Handlung nach Absatz 1
   a) Nummer 1 oder
   b) Nummer 2
   erlangt hat,
2. gegen eine Verpflichtung zur Beschränkung der Nutzung des Geschäftsgeheimnisses verstößt oder
3. gegen eine Verpflichtung verstößt, das Geschäftsgeheimnis nicht offenzulegen.

(3) [1] Ein Geschäftsgeheimnis darf nicht erlangen, nutzen oder offenlegen, wer das Geschäftsgeheimnis über eine andere Person erlangt hat und zum Zeitpunkt der Erlangung, Nutzung oder Offenlegung weiß oder wissen müsste, dass diese das Geschäftsgeheimnis entgegen Absatz 2 genutzt oder offengelegt hat. [2]Das gilt insbesondere, wenn die Nutzung in der Herstellung, dem Anbieten, dem Inverkehrbringen oder der Einfuhr, der Ausfuhr oder der Lagerung für diese Zwecke von rechtsverletzenden Produkten besteht.

§ 5 Ausnahmen

Die Erlangung, die Nutzung oder die Offenlegung eines Geschäftsgeheimnisses fällt nicht unter die Verbote des § 4, wenn dies zum Schutz eines berechtigten Interesses erfolgt, insbesondere

1. zur Ausübung des Rechts der freien Meinungsäußerung und der Informationsfreiheit, einschließlich der Achtung der Freiheit und der Pluralität der Medien;

2. zur Aufdeckung einer rechtswidrigen Handlung oder eines beruflichen oder sonstigen Fehlverhaltens, wenn die Erlangung, Nutzung oder Offenlegung geeignet ist, das allgemeine öffentliche Interesse zu schützen;
3. im Rahmen der Offenlegung durch Arbeitnehmer gegenüber der Arbeitnehmervertretung, wenn dies erforderlich ist, damit die Arbeitnehmervertretung ihre Aufgaben erfüllen kann.

Abschnitt 2 Ansprüche bei Rechtsverletzungen

§ 6 Beseitigung und Unterlassung

[1]Der Inhaber des Geschäftsgeheimnisses kann den Rechtsverletzer auf Beseitigung der Beeinträchtigung und bei Wiederholungsgefahr auch auf Unterlassung in Anspruch nehmen. [2]Der Anspruch auf Unterlassung besteht auch dann, wenn eine Rechtsverletzung erstmalig droht.

§ 7 Vernichtung; Herausgabe; Rückruf; Entfernung und Rücknahme vom Markt

Der Inhaber des Geschäftsgeheimnisses kann den Rechtsverletzer auch in Anspruch nehmen auf

1. Vernichtung oder Herausgabe der im Besitz oder Eigentum des Rechtsverletzers stehenden Dokumente, Gegenstände, Materialien, Stoffe oder elektronischen Dateien, die das Geschäftsgeheimnis enthalten oder verkörpern,
2. Rückruf des rechtsverletzenden Produkts,
3. dauerhafte Entfernung der rechtsverletzenden Produkte aus den Vertriebswegen,
4. Vernichtung der rechtsverletzenden Produkte oder
5. Rücknahme der rechtsverletzenden Produkte vom Markt, wenn der Schutz des Geschäftsgeheimnisses hierdurch nicht beeinträchtigt wird.

§ 8 Auskunft über rechtsverletzende Produkte; Schadensersatz bei Verletzung der Auskunftspflicht

(1) Der Inhaber des Geschäftsgeheimnisses kann vom Rechtsverletzer Auskunft über Folgendes verlangen:
   1. Name und Anschrift der Hersteller, Lieferanten und anderer Vorbesitzer der rechtsverletzenden Produkte sowie der gewerblichen Abnehmer und Verkaufsstellen, für die sie bestimmt waren,
   2. die Menge der hergestellten, bestellten, ausgelieferten oder erhaltenen rechtsverletzenden Produkte sowie über die Kaufpreise,
   3. diejenigen im Besitz oder Eigentum des Rechtsverletzers stehenden Dokumente, Gegenstände, Materialien, Stoffe oder elektronischen Dateien, die das Geschäftsgeheimnis enthalten oder verkörpern, und
   4. die Person, von der sie das Geschäftsgeheimnis erlangt haben und der gegenüber sie es offenbart haben.

(2) Erteilt der Rechtsverletzer vorsätzlich oder grob fahrlässig die Auskunft nicht, verspätet, falsch oder unvollständig, ist er dem Inhaber des Geschäftsgeheimnisses zum Ersatz des daraus entstehenden Schadens verpflichtet.

§ 9 Anspruchsausschluss bei Unverhältnismäßigkeit

Die Ansprüche nach den §§ 6 bis 8 Absatz 1 sind ausgeschlossen, wenn die Erfüllung im Einzelfall unverhältnismäßig wäre, unter Berücksichtigung insbesondere

1. des Wertes oder eines anderen spezifischen Merkmals des Geschäftsgeheimnisses,
2. der getroffenen Geheimhaltungsmaßnahmen,
3. des Verhaltens des Rechtsverletzers bei Erlangung, Nutzung oder Offenlegung des Geschäftsgeheimnisses,
4. der Folgen der rechtswidrigen Nutzung oder Offenlegung des Geschäftsgeheimnisses,
5. der berechtigten Interessen des Inhabers des Geschäftsgeheimnisses und des Rechtsverletzers sowie der Auswirkungen, die die Erfüllung der Ansprüche für beide haben könnte,
6. der berechtigten Interessen Dritter oder
7. des öffentlichen Interesses.

§ 10 Haftung des Rechtsverletzers

(1) [1] Ein Rechtsverletzer, der vorsätzlich oder fahrlässig handelt, ist dem Inhaber des Geschäftsgeheimnisses zum Ersatz des daraus entstehenden Schadens verpflichtet. [2]§ 619a des Bürgerlichen Gesetzbuchs bleibt unberührt.

(2) [1] Bei der Bemessung des Schadensersatzes kann auch der Gewinn, den der Rechtsverletzer durch die Verletzung des Rechts erzielt hat, berücksichtigt werden. [2]Der Schadensersatzanspruch kann auch auf der Grundlage des Betrages bestimmt werden, den der Rechtsverletzer als angemessene Vergütung hätte entrichten müssen, wenn er die Zustimmung zur Erlangung, Nutzung oder Offenlegung des Geschäftsgeheimnisses eingeholt hätte.

(3) Der Inhaber des Geschäftsgeheimnisses kann auch wegen des Schadens, der nicht Vermögensschaden ist, von dem Rechtsverletzer eine Entschädigung in Geld verlangen, soweit dies der Billigkeit entspricht.

§ 11 Abfindung in Geld

(1) Ein Rechtsverletzer, der weder vorsätzlich noch fahrlässig gehandelt hat, kann zur Abwendung der Ansprüche nach den §§ 6 oder 7 den Inhaber des Geschäftsgeheimnisses in Geld abfinden, wenn dem Rechtsverletzer durch die Erfüllung der Ansprüche ein unverhältnismäßig großer Nachteil entstehen würde und wenn die Abfindung in Geld als angemessen erscheint.

(2) [1] Die Höhe der Abfindung in Geld bemisst sich nach der Vergütung, die im Falle einer vertraglichen Einräumung des Nutzungsrechts angemessen wäre. [2]Sie darf den Betrag nicht übersteigen, der einer Vergütung im Sinne von Satz 1 für die Länge des Zeitraums entspricht, in dem dem Inhaber des Geschäftsgeheimnisses ein Unterlassungsanspruch zusteht.

[...]

§ 14 Missbrauchsverbot

[1]Die Geltendmachung der Ansprüche nach diesem Gesetz ist unzulässig, wenn sie unter Berücksichtigung der gesamten Umstände missbräuchlich ist. [2]Bei missbräuchlicher Geltendmachung kann der Anspruchsgegner Ersatz der für seine Rechtsverteidigung erforderlichen Aufwendungen verlangen. [3]Weitergehende Ersatzansprüche bleiben unberührt.

[...]

Abschnitt 4 Strafvorschriften

§ 23 Verletzung von Geschäftsgeheimnissen

(1) Mit Freiheitsstrafe bis zu drei Jahren oder mit Geldstrafe wird bestraft, wer zur Förderung des eigenen oder fremden Wettbewerbs, aus Eigennutz, zugunsten eines Dritten oder in der Absicht, dem Inhaber eines Unternehmens Schaden zuzufügen,
   1. entgegen § 4 Absatz 1 Nummer 1 ein Geschäftsgeheimnis erlangt,
   2. entgegen § 4 Absatz 2 Nummer 1 Buchstabe a ein Geschäftsgeheimnis nutzt oder offenlegt oder
   3. entgegen § 4 Absatz 2 Nummer 3 als eine bei einem Unternehmen beschäftigte Person ein Geschäftsgeheimnis, das ihr im Rahmen des Beschäftigungsverhältnisses anvertraut worden oder zugänglich geworden ist, während der Geltungsdauer des Beschäftigungsverhältnisses offenlegt.

(2) Ebenso wird bestraft, wer zur Förderung des eigenen oder fremden Wettbewerbs, aus Eigennutz, zugunsten eines Dritten oder in der Absicht, dem Inhaber eines Unternehmens Schaden zuzufügen, ein Geschäftsgeheimnis nutzt oder offenlegt, das er durch eine fremde Handlung nach Absatz 1 Nummer 2 oder Nummer 3 erlangt hat.

(3) Mit Freiheitsstrafe bis zu zwei Jahren oder mit Geldstrafe wird bestraft, wer zur Förderung des eigenen oder fremden Wettbewerbs oder aus Eigennutz entgegen § 4 Absatz 2 Nummer 2 oder Nummer 3 ein Geschäftsgeheimnis, das eine ihm im geschäftlichen Verkehr anvertraute geheime Vorlage oder Vorschrift technischer Art ist, nutzt oder offenlegt.

(4) Mit Freiheitsstrafe bis zu fünf Jahren oder mit Geldstrafe wird bestraft, wer
   1. in den Fällen des Absatzes 1 oder des Absatzes 2 gewerbsmäßig handelt,
   2. in den Fällen des Absatzes 1 Nummer 2 oder Nummer 3 oder des Absatzes 2 bei der Offenlegung weiß, dass das Geschäftsgeheimnis im Ausland genutzt werden soll, oder
   3. in den Fällen des Absatzes 1 Nummer 2 oder des Absatzes 2 das Geschäftsgeheimnis im Ausland nutzt.

(5) Der Versuch ist strafbar.

(6) Beihilfehandlungen einer in § 53 Absatz 1 Satz 1 Nummer 5 der Strafprozessordnung genannten Person sind nicht rechtswidrig, wenn sie sich auf die Entgegennahme, Auswertung oder Veröffentlichung des Geschäftsgeheimnisses beschränken.

(7) [1] § 5 Nummer 7 des Strafgesetzbuches gilt entsprechend. [2]Die §§ 30 und 31 des Strafgesetzbuches gelten entsprechend, wenn der Täter zur Förderung des eigenen oder fremden Wettbewerbs oder aus Eigennutz handelt.

(8) Die Tat wird nur auf Antrag verfolgt, es sei denn, dass die Strafverfolgungsbehörde wegen des besonderen öffentlichen Interesses an der Strafverfolgung ein Einschreiten von Amts wegen für geboten hält.

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)

[...]

in Erwägung nachstehender Gründe:
[...]

(63)
Eine betroffene Person sollte ein Auskunftsrecht hinsichtlich der sie betreffenden personenbezogenen Daten, die erhoben worden sind, besitzen und dieses Recht problemlos und in angemessenen Abständen wahrnehmen können, um sich der Verarbeitung bewusst zu sein und deren Rechtmäßigkeit überprüfen zu können. Dies schließt das Recht betroffene Personen auf Auskunft über ihre eigenen gesundheitsbezogenen Daten ein, etwa Daten in ihren Patientenakten, die Informationen wie beispielsweise Diagnosen, Untersuchungsergebnisse, Befunde der behandelnden Ärzte und Angaben zu Behandlungen oder Eingriffen enthalten. Jede betroffene Person sollte daher ein Anrecht darauf haben zu wissen und zu erfahren, insbesondere zu welchen Zwecken die personenbezogenen Daten verarbeitet werden und, wenn möglich, wie lange sie gespeichert werden, wer die Empfänger der personenbezogenen Daten sind, nach welcher Logik die automatische Verarbeitung personenbezogener Daten erfolgt und welche Folgen eine solche Verarbeitung haben kann, zumindest in Fällen, in denen die Verarbeitung auf Profiling beruht. Nach Möglichkeit sollte der Verantwortliche den Fernzugang zu einem sicheren System bereitstellen können, der der betroffenen Person direkten Zugang zu ihren personenbezogenen Daten ermöglichen würde. Dieses Recht sollte die Rechte und Freiheiten anderer Personen, etwa Geschäftsgeheimnisse oder Rechte des geistigen Eigentums und insbesondere das Urheberrecht an Software, nicht beeinträchtigen. Dies darf jedoch nicht dazu führen, dass der betroffenen Person jegliche Auskunft verweigert wird. Verarbeitet der Verantwortliche eine große Menge von Informationen

über die betroffene Person, so sollte er verlangen können, dass die betroffene Person präzisiert, auf welche Information oder welche Verarbeitungsvorgänge sich ihr Auskunftsersuchen bezieht, bevor er ihr Auskunft erteilt.

[...]

Kapitel II

Grundsätze

Artikel 5

Grundsätze für die Verarbeitung personenbezogener Daten

(1) Personenbezogene Daten müssen
   a) auf rechtmäßige Weise, nach Treu und Glauben und in einer für die betroffene Person nachvollziehbaren Weise verarbeitet werden ("Rechtmäßigkeit, Verarbeitung nach Treu und Glauben, Transparenz");
   b) für festgelegte, eindeutige und legitime Zwecke erhoben werden und dürfen nicht in einer mit diesen Zwecken nicht zu vereinbarenden Weise weiterverarbeitet werden; eine Weiterverarbeitung für im öffentlichen Interesse liegende Archivzwecke, für wissenschaftliche oder historische Forschungszwecke oder für statistische Zwecke gilt gemäß Artikel 89 Absatz 1 nicht als unvereinbar mit den ursprünglichen Zwecken ("Zweckbindung");
   c) dem Zweck angemessen und erheblich sowie auf das für die Zwecke der Verarbeitung notwendige Maß beschränkt sein ("Datenminimierung");
   d) sachlich richtig und erforderlichenfalls auf dem neuesten Stand sein; es sind alle angemessenen Maßnahmen zu treffen, damit personenbezogene Daten, die im Hinblick auf die Zwecke ihrer Verarbeitung unrichtig sind, unverzüglich gelöscht oder berichtigt werden ("Richtigkeit");
   e) in einer Form gespeichert werden, die die Identifizierung der betroffenen Personen nur so lange ermöglicht, wie es für die Zwecke, für die sie verarbeitet werden, erforderlich ist; personenbezogene Daten dürfen länger gespeichert werden, soweit die personenbezogenen Daten vorbehaltlich der Durchführung geeigneter technischer und organisatorischer Maßnahmen, die von dieser Verordnung zum Schutz der Rechte und Freiheiten der betroffenen Person gefordert werden, ausschließlich für

im öffentlichen Interesse liegende Archivzwecke oder für wissenschaftliche und historische Forschungszwecke oder für statistische Zwecke gemäß Artikel 89 Absatz 1 verarbeitet werden ("Speicherbegrenzung");

f) in einer Weise verarbeitet werden, die eine angemessene Sicherheit der personenbezogenen Daten gewährleistet, einschließlich Schutz vor unbefugter oder unrechtmäßiger Verarbeitung und vor unbeabsichtigtem Verlust, unbeabsichtigter Zerstörung oder unbeabsichtigter Schädigung durch geeignete technische und organisatorische Maßnahmen ("Integrität und Vertraulichkeit");

(2) Der Verantwortliche ist für die Einhaltung des Absatzes 1 verantwortlich und muss dessen Einhaltung nachweisen können ("Rechenschaftspflicht").

[...]

KAPITEL III

Rechte der betroffenen Person

Abschnitt 1

Transparenz und Modalitäten

Artikel 12

Transparente Information, Kommunikation und Modalitäten für die Ausübung der Rechte der betroffenen Person

(1) Der Verantwortliche trifft geeignete Maßnahmen, um der betroffenen Person alle Informationen gemäß den Artikeln 13 und 14 und alle Mitteilungen gemäß den Artikeln 15 bis 22 und Artikel 34, die sich auf die Verarbeitung beziehen, in präziser, transparenter, verständlicher und leicht zugänglicher Form in einer klaren und einfachen Sprache zu übermitteln; dies gilt insbesondere für Informationen, die sich speziell an Kinder richten. Die Übermittlung der Informationen erfolgt schriftlich oder in anderer Form, gegebenenfalls auch elektronisch. Falls von der betroffenen Person verlangt, kann die Information mündlich erteilt werden, sofern die Identität der betroffenen Person in anderer Form nachgewiesen wurde.

(2) Der Verantwortliche erleichtert der betroffenen Person die Ausübung ihrer Rechte gemäß den Artikeln 15 bis 22. In den in Artikel 11 Absatz 2 genannten Fällen darf sich der Verantwortliche nur dann weigern, aufgrund des Antrags der betroffenen Person auf Wahrnehmung ihrer Rechte gemäß den Artikeln 15 bis 22 tätig zu werden, wenn er glaubhaft macht, dass er nicht in der Lage ist, die betroffene Person zu identifizieren.

(3) Der Verantwortliche stellt der betroffenen Person Informationen über die auf Antrag gemäß den Artikeln 15 bis 22 ergriffenen Maßnahmen unverzüglich, in jedem Fall aber innerhalb eines Monats nach Eingang des Antrags zur Verfügung. Diese Frist kann um weitere zwei Monate verlängert werden, wenn dies unter Berücksichtigung der Komplexität und der Anzahl von Anträgen erforderlich ist. Der Verantwortliche unterrichtet die betroffene Person innerhalb eines Monats nach Eingang des Antrags über eine Fristverlängerung, zusammen mit den Gründen für die Verzögerung. Stellt die betroffene Person den Antrag elektronisch, so ist sie nach Möglichkeit auf elektronischem Weg zu unterrichten, sofern sie nichts anderes angibt.

(4) Wird der Verantwortliche auf den Antrag der betroffenen Person hin nicht tätig, so unterrichtet er die betroffene Person ohne Verzögerung, spätestens aber innerhalb eines Monats nach Eingang des Antrags über die Gründe hierfür und über die Möglichkeit, bei einer Aufsichtsbehörde Beschwerde einzulegen oder einen gerichtlichen Rechtsbehelf einzulegen.

(5) Informationen gemäß den Artikeln 13 und 14 sowie alle Mitteilungen und Maßnahmen gemäß den Artikeln 15 bis 22 und Artikel 34 werden unentgeltlich zur Verfügung gestellt. Bei offenkundig unbegründeten oder - insbesondere im Fall von häufiger Wiederholung - exzessiven Anträgen einer betroffenen Person kann der Verantwortliche entweder
   a) ein angemessenes Entgelt verlangen, bei dem die Verwaltungskosten für die Unterrichtung oder die Mitteilung oder die Durchführung der beantragten Maßnahme berücksichtigt werden, oder
   b) sich weigern, aufgrund des Antrags tätig zu werden.

   Der Verantwortliche hat den Nachweis für den offenkundig unbegründeten oder exzessiven Charakter des Antrags zu erbringen.

(6) Hat der Verantwortliche begründete Zweifel an der Identität der natürlichen Person, die den Antrag gemäß den Artikeln 15 bis 21 stellt, so kann er

unbeschadet des Artikels 11 zusätzliche Informationen anfordern, die zur Bestätigung der Identität der betroffenen Person erforderlich sind.

(7) Die Informationen, die den betroffenen Personen gemäß den Artikeln 13 und 14 bereitzustellen sind, können in Kombination mit standardisierten Bildsymbolen bereitgestellt werden, um in leicht wahrnehmbarer, verständlicher und klar nachvollziehbarer Form einen aussagekräftigen Überblick über die beabsichtigte Verarbeitung zu vermitteln. Werden die Bildsymbole in elektronischer Form dargestellt, müssen sie maschinenlesbar sein.

(8) Der Kommission wird die Befugnis übertragen, gemäß Artikel 92 delegierte Rechtsakte zur Bestimmung der Informationen, die durch Bildsymbole darzustellen sind, und der Verfahren für die Bereitstellung standardisierter Bildsymbole zu erlassen.

Abschnitt 2

Informationspflicht und Recht auf Auskunft zu personenbezogenen Daten

Artikel 13

Informationspflicht bei Erhebung von personenbezogenen Daten bei der betroffenen Person

(1) Werden personenbezogene Daten bei der betroffenen Person erhoben, so teilt der Verantwortliche der betroffenen Person zum Zeitpunkt der Erhebung dieser Daten Folgendes mit:

a) den Namen und die Kontaktdaten des Verantwortlichen sowie gegebenenfalls seines Vertreters;

b) gegebenenfalls die Kontaktdaten des Datenschutzbeauftragten;

c) die Zwecke, für die die personenbezogenen Daten verarbeitet werden sollen, sowie die Rechtsgrundlage für die Verarbeitung;

d) wenn die Verarbeitung auf Artikel 6 Absatz 1 Buchstabe f beruht, die berechtigten Interessen, die von dem Verantwortlichen oder einem Dritten verfolgt werden;

e) gegebenenfalls die Empfänger oder Kategorien von Empfängern der personenbezogenen Daten und

f) gegebenenfalls die Absicht des Verantwortlichen, die personenbezogenen Daten an ein Drittland oder eine internationale Organisation zu übermitteln, sowie das Vorhandensein oder das Fehlen eines Angemessenheitsbeschlusses der Kommission oder im Falle von Übermittlungen gemäß Artikel 46 oder Artikel 47 oder Artikel 49 Absatz 1 Unterabsatz 2 einen Verweis auf die geeigneten oder angemessenen Garantien und die Möglichkeit, wie eine Kopie von ihnen zu erhalten ist, oder wo sie verfügbar sind.

(2) Zusätzlich zu den Informationen gemäß Absatz 1 stellt der Verantwortliche der betroffenen Person zum Zeitpunkt der Erhebung dieser Daten folgende weitere Informationen zur Verfügung, die notwendig sind, um eine faire und transparente Verarbeitung zu gewährleisten:

a) die Dauer, für die die personenbezogenen Daten gespeichert werden oder, falls dies nicht möglich ist, die Kriterien für die Festlegung dieser Dauer;

b) das Bestehen eines Rechts auf Auskunft seitens des Verantwortlichen über die betreffenden personenbezogenen Daten sowie auf Berichtigung oder Löschung oder auf Einschränkung der Verarbeitung oder eines Widerspruchsrechts gegen die Verarbeitung sowie des Rechts auf Datenübertragbarkeit;

c) wenn die Verarbeitung auf Artikel 6 Absatz 1 Buchstabe a oder Artikel 9 Absatz 2 Buchstabe a beruht, das Bestehen eines Rechts, die Einwilligung jederzeit zu widerrufen, ohne dass die Rechtmäßigkeit der aufgrund der Einwilligung bis zum Widerruf erfolgten Verarbeitung berührt wird;

d) das Bestehen eines Beschwerderechts bei einer Aufsichtsbehörde;

e) ob die Bereitstellung der personenbezogenen Daten gesetzlich oder vertraglich vorgeschrieben oder für einen Vertragsabschluss erforderlich ist, ob die betroffene Person verpflichtet ist, die personenbezogenen Daten bereitzustellen, und welche mögliche Folgen die Nichtbereitstellung hätte und

f) das Bestehen einer automatisierten Entscheidungsfindung einschließlich Profiling gemäß Artikel 22 Absätze 1 und 4 und – zumindest in diesen Fällen – aussagekräftige Informationen über die involvierte Logik sowie die Tragweite und die angestrebten Auswirkungen einer derartigen Verarbeitung für die betroffene Person.

(3) Beabsichtigt der Verantwortliche, die personenbezogenen Daten für einen anderen Zweck weiterzuverarbeiten als den, für den die personenbezogenen Daten erhoben wurden, so stellt er der betroffenen Person vor dieser Weiterverarbeitung Informationen über diesen anderen Zweck und alle anderen maßgeblichen Informationen gemäß Absatz 2 zur Verfügung.
(4) Die Absätze 1, 2 und 3 finden keine Anwendung, wenn und soweit die betroffene Person bereits über die Informationen verfügt.

Artikel 14

Informationspflicht, wenn die personenbezogenen Daten nicht bei der betroffenen Person erhoben wurden

(1) Werden personenbezogene Daten nicht bei der betroffenen Person erhoben, so teilt der Verantwortliche der betroffenen Person Folgendes mit:
   a) den Namen und die Kontaktdaten des Verantwortlichen sowie gegebenenfalls seines Vertreters;
   b) zusätzlich die Kontaktdaten des Datenschutzbeauftragten;
   c) die Zwecke, für die die personenbezogenen Daten verarbeitet werden sollen, sowie die Rechtsgrundlage für die Verarbeitung;
   d) die Kategorien personenbezogener Daten, die verarbeitet werden;
   e) gegebenenfalls die Empfänger oder Kategorien von Empfängern der personenbezogenen Daten;
   f) gegebenenfalls die Absicht des Verantwortlichen, die personenbezogenen Daten an einen Empfänger in einem Drittland oder einer internationalen Organisation zu übermitteln, sowie das Vorhandensein oder das Fehlen eines Angemessenheitsbeschlusses der Kommission oder im Falle von Übermittlungen gemäß Artikel 46 oder Artikel 47 oder Artikel 49 Absatz 1 Unterabsatz 2 einen Verweis auf die geeigneten oder angemessenen Garantien und die Möglichkeit, eine Kopie von ihnen zu erhalten, oder wo sie verfügbar sind.
(2) Zusätzlich zu den Informationen gemäß Absatz 1 stellt der Verantwortliche der betroffenen Person die folgenden Informationen zur Verfügung, die erforderlich sind, um der betroffenen Person gegenüber eine faire und transparente Verarbeitung zu gewährleisten:

a) die Dauer, für die die personenbezogenen Daten gespeichert werden oder, falls dies nicht möglich ist, die Kriterien für die Festlegung dieser Dauer;
b) wenn die Verarbeitung auf Artikel 6 Absatz 1 Buchstabe f beruht, die berechtigten Interessen, die von dem Verantwortlichen oder einem Dritten verfolgt werden;
c) das Bestehen eines Rechts auf Auskunft seitens des Verantwortlichen über die betreffenden personenbezogenen Daten sowie auf Berichtigung oder Löschung oder auf Einschränkung der Verarbeitung und eines Widerspruchsrechts gegen die Verarbeitung sowie des Rechts auf Datenübertragbarkeit;
d) wenn die Verarbeitung auf Artikel 6 Absatz 1 Buchstabe a oder Artikel 9 Absatz 2 Buchstabe a beruht, das Bestehen eines Rechts, die Einwilligung jederzeit zu widerrufen, ohne dass die Rechtmäßigkeit der aufgrund der Einwilligung bis zum Widerruf erfolgten Verarbeitung berührt wird;
e) das Bestehen eines Beschwerderechts bei einer Aufsichtsbehörde;
f) aus welcher Quelle die personenbezogenen Daten stammen und gegebenenfalls ob sie aus öffentlich zugänglichen Quellen stammen;
g) das Bestehen einer automatisierten Entscheidungsfindung einschließlich Profiling gemäß Artikel 22 Absätze 1 und 4 und – zumindest in diesen Fällen – aussagekräftige Informationen über die involvierte Logik sowie die Tragweite und die angestrebten Auswirkungen einer derartigen Verarbeitung für die betroffene Person.

(3) Der Verantwortliche erteilt die Informationen gemäß den Absätzen 1 und 2
a) unter Berücksichtigung der spezifischen Umstände der Verarbeitung der personenbezogenen Daten innerhalb einer angemessenen Frist nach Erlangung der personenbezogenen Daten, längstens jedoch innerhalb eines Monats,
b) falls die personenbezogenen Daten zur Kommunikation mit der betroffenen Person verwendet werden sollen, spätestens zum Zeitpunkt der ersten Mitteilung an sie, oder,
c) falls die Offenlegung an einen anderen Empfänger beabsichtigt ist, spätestens zum Zeitpunkt der ersten Offenlegung.

(4) Beabsichtigt der Verantwortliche, die personenbezogenen Daten für einen anderen Zweck weiterzuverarbeiten als den, für den die personenbezoge-

nen Daten erlangt wurden, so stellt er der betroffenen Person vor dieser Weiterverarbeitung Informationen über diesen anderen Zweck und alle anderen maßgeblichen Informationen gemäß Absatz 2 zur Verfügung.

(5) Die Absätze 1 bis 4 finden keine Anwendung, wenn und soweit
   a) die betroffene Person bereits über die Informationen verfügt,
   b) die Erteilung dieser Informationen sich als unmöglich erweist oder einen unverhältnismäßigen Aufwand erfordern würde; dies gilt insbesondere für die Verarbeitung für im öffentlichen Interesse liegende Archivzwecke, für wissenschaftliche oder historische Forschungszwecke oder für statistische Zwecke vorbehaltlich der in Artikel 89 Absatz 1 genannten Bedingungen und Garantien oder soweit die in Absatz 1 des vorliegenden Artikels genannte Pflicht voraussichtlich die Verwirklichung der Ziele dieser Verarbeitung unmöglich macht oder ernsthaft beeinträchtigt In diesen Fällen ergreift der Verantwortliche geeignete Maßnahmen zum Schutz der Rechte und Freiheiten sowie der berechtigten Interessen der betroffenen Person, einschließlich der Bereitstellung dieser Informationen für die Öffentlichkeit,
   c) die Erlangung oder Offenlegung durch Rechtsvorschriften der Union oder der Mitgliedstaaten, denen der Verantwortliche unterliegt und die geeignete Maßnahmen zum Schutz der berechtigten Interessen der betroffenen Person vorsehen, ausdrücklich geregelt ist oder
   d) die personenbezogenen Daten gemäß dem Unionsrecht oder dem Recht der Mitgliedstaaten dem Berufsgeheimnis, einschließlich einer satzungsmäßigen Geheimhaltungspflicht, unterliegen und daher vertraulich behandelt werden müssen.

Artikel 15

Auskunftsrecht der betroffenen Person

(1) Die betroffene Person hat das Recht, von dem Verantwortlichen eine Bestätigung darüber zu verlangen, ob sie betreffende personenbezogene Daten verarbeitet werden; ist dies der Fall, so hat sie ein Recht auf Auskunft über diese personenbezogenen Daten und auf folgende Informationen:
   a) die Verarbeitungszwecke;
   b) die Kategorien personenbezogener Daten, die verarbeitet werden;

c) die Empfänger oder Kategorien von Empfängern, gegenüber denen die personenbezogenen Daten offengelegt worden sind oder noch offengelegt werden, insbesondere bei Empfängern in Drittländern oder bei internationalen Organisationen;
d) falls möglich die geplante Dauer, für die die personenbezogenen Daten gespeichert werden, oder, falls dies nicht möglich ist, die Kriterien für die Festlegung dieser Dauer;
e) das Bestehen eines Rechts auf Berichtigung oder Löschung der sie betreffenden personenbezogenen Daten oder auf Einschränkung der Verarbeitung durch den Verantwortlichen oder eines Widerspruchsrechts gegen diese Verarbeitung;
f) das Bestehen eines Beschwerderechts bei einer Aufsichtsbehörde;
g) wenn die personenbezogenen Daten nicht bei der betroffenen Person erhoben werden, alle verfügbaren Informationen über die Herkunft der Daten;
h) das Bestehen einer automatisierten Entscheidungsfindung einschließlich Profiling gemäß Artikel 22 Absätze 1 und 4 und – zumindest in diesen Fällen – aussagekräftige Informationen über die involvierte Logik sowie die Tragweite und die angestrebten Auswirkungen einer derartigen Verarbeitung für die betroffene Person.

(2) Werden personenbezogene Daten an ein Drittland oder an eine internationale Organisation übermittelt, so hat die betroffene Person das Recht, über die geeigneten Garantien gemäß Artikel 46 im Zusammenhang mit der Übermittlung unterrichtet zu werden.

(3) Der Verantwortliche stellt eine Kopie der personenbezogenen Daten, die Gegenstand der Verarbeitung sind, zur Verfügung. Für alle weiteren Kopien, die die betroffene Person beantragt, kann der Verantwortliche ein angemessenes Entgelt auf der Grundlage der Verwaltungskosten verlangen. Stellt die betroffene Person den Antrag elektronisch, so sind die Informationen in einem gängigen elektronischen Format zur Verfügung zu stellen, sofern sie nichts anderes angibt.

(4) Das Recht auf Erhalt einer Kopie gemäß Absatz 3 darf die Rechte und Freiheiten anderer Personen nicht beeinträchtigen.

[...]

Artikel 22

Automatisierte Entscheidungen im Einzelfall einschließlich Profiling

(1) Die betroffene Person hat das Recht, 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.

(2) Absatz 1 gilt nicht, wenn die Entscheidung
   a) für den Abschluss oder die Erfüllung eines Vertrags zwischen der betroffenen Person und dem Verantwortlichen erforderlich ist,
   b) aufgrund von Rechtsvorschriften der Union oder der Mitgliedstaaten, denen der Verantwortliche unterliegt, zulässig ist und diese Rechtsvorschriften angemessene Maßnahmen zur Wahrung der Rechte und Freiheiten sowie der berechtigten Interessen der betroffenen Person enthalten oder
   c) mit ausdrücklicher Einwilligung der betroffenen Person erfolgt.

(3) In den in Absatz 2 Buchstaben a und c genannten Fällen trifft der Verantwortliche angemessene Maßnahmen, um die Rechte und Freiheiten sowie die berechtigten Interessen der betroffenen Person zu wahren, wozu mindestens das Recht auf Erwirkung des Eingreifens einer Person seitens des Verantwortlichen, auf Darlegung des eigenen Standpunkts und auf Anfechtung der Entscheidung gehört.

(4) Entscheidungen nach Absatz 2 dürfen nicht auf besonderen Kategorien personenbezogener Daten nach Artikel 9 Absatz 1 beruhen, sofern nicht Artikel 9 Absatz 2 Buchstabe a oder g gilt und angemessene Maßnahmen zum Schutz der Rechte und Freiheiten sowie der berechtigten Interessen der betroffenen Person getroffen wurden.

Vorschlag für eine

VERORDNUNG DES EUROPÄISCHEN PARLAMENTS UND DES RATES ZUR FESTLEGUNG HARMONISIERTER VORSCHRIFTEN FÜR KÜNSTLICHE INTELLIGENZ (GESETZ ÜBER KÜNSTLICHE INTELLIGENZ) UND ZUR ÄNDERUNG BESTIMMTER RECHTSAKTE DER UNION

Allgemeine Ausrichtung vom 6.12.2022, 15698/22

DAS EUROPÄISCHE PARLAMENT UND DER RAT
DER EUROPÄISCHEN UNION

[...]

in Erwägung nachstehender Gründe:
[...]

(64) Angesichts der umfassenderen Erfahrung professioneller dem Inverkehrbringen vorgeschalteter Zertifizierer im Bereich der Produktsicherheit und der unterschiedlichen Art der damit verbundenen Risiken empfiehlt es sich, zumindest während der anfänglichen Anwendung dieser Verordnung für Hochrisiko-KI-Systeme, die nicht mit Produkten in Verbindung stehen, den Anwendungsbereich der Konformitätsbewertung durch Dritte einzuschränken. Daher sollte die Konformitätsbewertung solcher Systeme in der Regel vom Anbieter in eigener Verantwortung durchgeführt werden, mit Ausnahme von KI-Systemen, die zur biometrischen Fernidentifizierung von Personen verwendet werden sollen, bei denen die Beteiligung einer notifizierten Stelle an der Konformitätsbewertung vorgesehen werden sollte, soweit diese Systeme nicht ganz verboten sind.

[...]

(83) Zur Gewährleistung einer vertrauensvollen und konstruktiven Zusammenarbeit der zuständigen Behörden auf Ebene der Union und der Mitglied-

staaten sollten alle an der Anwendung dieser Verordnung beteiligten Parteien im Einklang mit dem Unionsrecht und dem nationalen Recht die Vertraulichkeit der im Rahmen der Durchführung ihrer Tätigkeiten erlangten Informationen und Daten wahren.

[...]

HABEN FOLGENDE VERORDNUNG ERLASSEN:

TITEL I

ALLGEMEINE BESTIMMUNGEN

Artikel 1

Gegenstand

In dieser Verordnung wird Folgendes festgelegt:

a) harmonisierte Vorschriften für das Inverkehrbringen, die Inbetriebnahme und die Verwendung von Systemen der künstlichen Intelligenz (im Folgenden „KI-Systeme“) in der Union;
a) Verbote bestimmter Praktiken im Bereich der künstlichen Intelligenz;
b) besondere Anforderungen an Hochrisiko-KI-Systeme und Verpflichtungen für Akteure in Bezug auf solche Systeme;
c) harmonisierte Transparenzvorschriften für bestimmte KI-Systeme;
d) Vorschriften für die Marktbeobachtung, Marktüberwachung und Governance;
e) Maßnahmen zur Innovationsförderung.

Artikel 2

Anwendungsbereich

(1) Diese Verordnung gilt für:
   a) Anbieter, die KI-Systeme in der Union in Verkehr bringen oder in Betrieb nehmen, unabhängig davon, ob diese Anbieter in der Union oder in einem Drittland physisch anwesend oder niedergelassen sind;
   b) Nutzer von KI-Systemen, die in der Union physisch anwesend oder niedergelassen sind;
   c) Anbieter und Nutzer von KI-Systemen, die in einem Drittland physisch anwesend oder niedergelassen sind, wenn das vom System hervorgebrachte Ergebnis in der Union verwendet wird;
   d) Einführer und Händler von KI-Systemen;
   e) Produkthersteller, die KI-Systeme zusammen mit ihrem Produkt unter ihrem Namen oder ihrer Handelsmarke in Verkehr bringen oder in Betrieb nehmen;
   f) Bevollmächtigte von Anbietern, die in der Union niedergelassen sind.

(2) Für KI-Systeme, die als Hochrisiko-KI-Systeme gemäß Artikel 6 Absätze 1 und 2 eingestuft sind und sich auf Produkte beziehen, die unter die in Anhang II Abschnitt B aufgeführten Harmonisierungsrechtsvorschriften der Union fallen, gilt nur Artikel 84 dieser Verordnung. Artikel 53 gilt nur, soweit die Anforderungen an Hochrisiko-KI-Systeme gemäß dieser Verordnung im Rahmen der genannten Harmonisierungsrechtsvorschriften der Union eingebunden wurden.

(3) Diese Verordnung gilt nicht für KI-Systeme, wenn und soweit sie mit oder ohne Änderungen für die Zwecke von Tätigkeiten, die nicht in den Anwendungsbereich des Unionsrechts fallen, in Verkehr gebracht, in Betrieb genommen oder verwendet werden, und in keinem Fall für Tätigkeiten in Bezug auf militärische Angelegenheiten, Verteidigung und nationale Sicherheit, unabhängig von der Art der Einrichtung, die diese Tätigkeiten ausübt.

   Darüber hinaus gilt diese Verordnung nicht für KI-Systeme, die nicht in der Union in Verkehr gebracht oder in Betrieb genommen werden, wenn die Ergebnisse in der Union für die Zwecke von Tätigkeiten, die nicht in den Anwendungsbereich des Unionsrechts fallen, verwendet werden, und in keinem Fall für Tätigkeiten in Bezug auf militärische Angelegenheiten,

Verteidigung und nationale Sicherheit, unabhängig von der Art der Einrichtung, die diese Tätigkeiten ausübt.

(4) Diese Verordnung gilt weder für Behörden in Drittländern noch für internationale Organisationen, die gemäß Absatz 1 in den Anwendungsbereich dieser Verordnung fallen, soweit diese Behörden oder Organisationen KI-Systeme im Rahmen internationaler Übereinkünfte im Bereich der Strafverfolgung und justiziellen Zusammenarbeit mit der Union oder mit einem oder mehreren Mitgliedstaaten verwenden.

(5) Die Anwendung der Bestimmungen über die Verantwortlichkeit der Vermittler in Kapitel II Abschnitt 4 der Richtlinie 2000/31/EG des Europäischen Parlaments und des Rates [*die durch die entsprechenden Bestimmungen des Gesetzes über digitale Dienste ersetzt werden sollen*] bleibt von dieser Verordnung unberührt.

(6) Diese Verordnung gilt nicht für KI-Systeme und deren Ergebnisse, die eigens für den alleinigen Zweck der wissenschaftlichen Forschung und Entwicklung entwickelt und in Betrieb genommen werden.

(7) Diese Verordnung gilt nicht für Forschungs- und Entwicklungsaktivitäten zu KI-Systemen.

(8) Diese Verordnung gilt nicht für die Pflichten von Nutzern, die natürliche Personen sind und KI-Systeme im Rahmen einer ausschließlich persönlichen und nicht beruflichen Tätigkeit verwenden, mit Ausnahme von Artikel 52.

Artikel 3

Begriffsbestimmungen

Für die Zwecke dieser Verordnung bezeichnet der Ausdruck

1. „System der künstlichen Intelligenz“ (KI-System) ein System, das so konzipiert ist, dass es mit Elementen der Autonomie arbeitet, und das auf der Grundlage maschineller und/oder vom Menschen erzeugter Daten und Eingaben durch maschinelles Lernen und/oder logik- und wissensgestützte Konzepte ableitet, wie eine Reihe von Zielen erreicht wird, und systemgenerierte Ergebnisse wie Inhalte (generative KI-Systeme), Vorhersagen, Empfehlungen oder Entscheidungen hervorbringt, die das Umfeld beeinflussen, mit dem die KI-Systeme interagieren;

[...]

2. „Anbieter“ eine natürliche oder juristische Person, Behörde, Einrichtung oder sonstige Stelle, die ein KI-System entwickelt oder entwickeln lässt und dieses System unter dem eigenen Namen oder der eigenen Marke in Verkehr bringt oder in Betrieb nimmt, sei es entgeltlich oder unentgeltlich;

[...]

4. „Nutzer“ eine natürliche oder juristische Person, einschließlich Behörden, Einrichtungen oder sonstige Stellen, unter deren Verantwortung das System verwendet wird;

[...]

20. „notifizierende Behörde“ die nationale Behörde, die für die Einrichtung und Durchführung der erforderlichen Verfahren für die Bewertung, Benennung und Notifizierung von Konformitätsbewertungsstellen und für deren Überwachung zuständig ist;
21. „Konformitätsbewertungsstelle“ eine Stelle, die Konformitätsbewertungstätigkeiten einschließlich Prüfungen, Zertifizierungen und Kontrollen durchführt und dabei als unabhängige Dritte auftritt;
22. „notifizierte Stelle“ eine Konformitätsbewertungsstelle, die gemäß dieser Verordnung und anderen einschlägigen Harmonisierungsvorschriften der Union benannt wurde;

[...]

26. „Marktüberwachungsbehörde“ die nationale Behörde, die die Tätigkeiten durchführt und die Maßnahmen ergreift, die in der Verordnung (EU) 2019/1020 vorgesehen sind;

[...]

43. „zuständige nationale Behörde“ die folgenden Behörden: die notifizierende Behörde und die Marktüberwachungsbehörde. In Bezug auf KI-Systeme, die von Organen, Einrichtungen und sonstigen Stellen der EU in Betrieb genommen oder verwendet werden, übernimmt der Europäische Datenschutzbeauftragte die Zuständigkeiten, die in den Mitgliedstaaten den zuständigen nationalen Behörden zugewiesen werden, und jede Bezugnahme auf die zuständigen nationalen Behörden oder Marktüberwachungsbehörden in dieser Verordnung ist gegebenenfalls als Bezugnahme auf den Europäischen Datenschutzbeauftragten zu verstehen;

[...]

Artikel 4b

Anforderungen an KI-Systeme mit allgemeinem Verwendungszweck und Pflichten der Anbieter solcher Systeme

(1) KI-Systeme mit allgemeinem Verwendungszweck, die als Hochrisiko-KI-Systeme oder als Komponenten von Hochrisiko-KI-Systemen im Sinne von Artikel 6 verwendet werden können, erfüllen die in Titel III Kapitel 2 dieser Verordnung festgelegten Anforderungen ab dem Datum der Anwendung der Durchführungsrechtsakte, die von der Kommission im Einklang mit dem in Artikel 74 Absatz 2 genannten Prüfverfahren erlassen werden, spätestens jedoch 18 Monate nach Inkrafttreten dieser Verordnung. In diesen Durchführungsrechtsakten wird die Anwendung der in Titel III Kapitel 2 festgelegten Anforderungen präzisiert und an KI-Systeme mit allgemeinem Verwendungszweck angepasst, und zwar im Hinblick auf ihre Merkmale, die technische Durchführbarkeit, die Besonderheiten der KI-Wertschöpfungskette sowie die Marktentwicklungen und technischen Entwicklungen. Bei der Erfüllung dieser Anforderungen wird dem allgemein anerkannten Stand der Technik Rechnung getragen.

(2) Anbieter von KI-Systemen mit allgemeinem Verwendungszweck nach Absatz 1 erfüllen die in den Artikeln 16aa, 16e, 16f, 16g, 16i, 16j, 25, 48 und 61 festgelegten Verpflichtungen ab dem Datum der Anwendung der in Absatz 1 genannten Durchführungsrechtsakte.

(3) Für die Zwecke der Erfüllung der Verpflichtungen nach Artikel 16e wenden Anbieter das in Anhang VI Nummern 3 und 4 festgelegte Konformitätsbewertungsverfahren auf der Grundlage einer internen Kontrolle an.

(4) Anbieter solcher Systeme halten die in Artikel 11 genannte technische Dokumentation für einen Zeitraum von zehn Jahren ab dem Inverkehrbringen oder der Inbetriebnahme des KI-Systems mit allgemeinem Verwendungszweck in der Union für die zuständigen nationalen Behörden bereit.

(5) Anbieter von KI-Systemen mit allgemeinem Verwendungszweck arbeiten mit anderen Anbietern zusammen, die beabsichtigen, solche Systeme als Hochrisiko-KI-Systeme oder als Komponenten von Hochrisiko-KI-Systemen in der Union in Betrieb zu nehmen oder in Verkehr zu bringen, und stellen ihnen die erforderlichen Informationen zur Verfügung, damit sie ihren Verpflichtungen aus dieser Verordnung nachkommen können. Bei dieser Zusammenarbeit zwischen Anbietern werden gegebenenfalls die Rechte des geistigen Eigentums sowie Betriebs- oder Geschäftsgeheimnisse

gemäß Artikel 70 gewahrt. Zur Gewährleistung einheitlicher Bedingungen für die Umsetzung dieser Verordnung in Bezug auf den Austausch von Informationen zwischen Anbietern von KI-Systemen mit allgemeinem Verwendungszweck, kann die Kommission gemäß dem in Artikel 74 Absatz 2 genannten Prüfverfahren Durchführungsrechtsakte erlassen.

(6) Bei der Erfüllung der in den Absätzen 1, 2 und 3 genannten Anforderungen und Verpflichtungen
   - ist jede Bezugnahme auf die Zweckbestimmung als Bezugnahme auf die mögliche Verwendung von KI-Systemen mit allgemeinem Verwendungszweck als Hochrisiko-KI-Systeme oder als Komponenten von Hochrisiko-KI-Systemen im Sinne von Artikel 6 zu verstehen;
   - ist jede Bezugnahme auf die Anforderungen an Hochrisiko-KI-Systeme in Titel III Kapitel 2 so zu verstehen, dass sie sich nur auf die in diesem Artikel festgelegten Anforderungen bezieht.

[...]

Artikel 11

Technische Dokumentation

(1) Die technische Dokumentation eines Hochrisiko-KI-Systems wird erstellt, bevor dieses System in Verkehr gebracht oder in Betrieb genommen wird, und ist stets auf dem neuesten Stand zu halten.
Die technische Dokumentation wird so erstellt, dass aus ihr der Nachweis hervorgeht, wie das Hochrisiko-KI-System die Anforderungen dieses Kapitels erfüllt, und dass den zuständigen nationalen Behörden und den notifizierten Stellen alle Informationen in klarer und verständlicher Form zur Verfügung stehen, die erforderlich sind, um zu beurteilen, ob das KI-System diese Anforderungen erfüllt. Sie enthält zumindest die in Anhang IV genannten Angaben oder – im Falle von KMU und Start-up-Unternehmen – alle gleichwertigen Unterlagen, die denselben Zwecken dienen, sofern die zuständige Behörde dies nicht als unangemessen erachtet.

(2) Wird ein Hochrisiko-KI-System, das mit einem Produkt verbunden ist, das unter die in Anhang II Abschnitt A aufgeführten Rechtsakte fällt, in Verkehr gebracht oder in Betrieb genommen, so wird eine einzige technische

Dokumentation erstellt, die alle in Anhang IV genannten Informationen sowie die nach diesen Rechtsakten erforderlichen Informationen enthält.

(3) Der Kommission wird die Befugnis übertragen, gemäß Artikel 73 delegierte Rechtsakte zur Änderung des Anhangs IV zu erlassen, wenn dies nötig ist, damit die technische Dokumentation in Anbetracht des technischen Fortschritts stets alle Informationen enthält, die erforderlich sind, um zu beurteilen, ob das System die Anforderungen dieses Kapitels erfüllt.

[...]

Artikel 13

Transparenz und Bereitstellung von Informationen für die Nutzer

(1) Hochrisiko-KI-Systeme werden so konzipiert und entwickelt, dass ihr Betrieb hinreichend transparent ist, damit die Nutzer und Anbieter ihre in Kapitel 3 dieses Titels festgelegten einschlägigen Pflichten erfüllen können und damit die Nutzer das System angemessen verstehen und verwenden können.

(2) Hochrisiko-KI-Systeme werden mit Gebrauchsanweisungen in einem geeigneten digitalen Format bereitgestellt oder auf andere Weise mit Gebrauchsanweisungen versehen, die präzise, vollständige, korrekte und eindeutige Informationen in einer für die Nutzer relevanten, barrierefrei zugänglichen und verständlichen Form enthalten.

(3) Die in Absatz 2 genannten Informationen umfassen:
   a) den Namen und die Kontaktangaben des Anbieters sowie gegebenenfalls seines Bevollmächtigten;
   b) die Merkmale, Fähigkeiten und Leistungsgrenzen des Hochrisiko-KI-Systems, einschließlich
      i) seiner Zweckbestimmung, auch der besonderen geografischen, verhaltensbezogenen oder funktionalen Rahmenbedingungen, unter denen ein Hochrisiko-KI-System bestimmungsgemäß verwendet werden soll,
      ii) des Genauigkeitsgrads – auch seiner Kennzahlen –, Robustheit und Cybersicherheit gemäß Artikel 15, für das das Hochrisiko-KI-System getestet und validiert wurde und das zu erwarten ist, sowie alle bekannten und vorhersehbaren Umstände, die sich auf das er-

wartete Maß an Genauigkeit, Robustheit und Cybersicherheit auswirken können,

iii) aller bekannten oder vorhersehbaren Umstände im Zusammenhang mit der bestimmungsgemäßen Verwendung des Hochrisiko-KI-Systems, die zu den in Artikel 9 Absatz 2 genannten Risiken für die Gesundheit und Sicherheit, die Grundrechte oder die Umwelt führen können,

iv) gegebenenfalls seines Verhaltens gegenüber bestimmten Personen oder Personengruppen, auf die das System bestimmungsgemäß angewandt werden soll,

v) gegebenenfalls der Spezifikationen für die Eingabedaten oder sonstiger relevanter Informationen über die verwendeten Trainings-, Validierungs- und Testdatensätze unter Berücksichtigung der Zweckbestimmung des KI-Systems;

vi) gegebenenfalls der Beschreibung des erwarteten Ergebnisses des Systems.

c) etwaige Änderungen des Hochrisiko-KI-Systems und seiner Leistung, die der Anbieter zum Zeitpunkt der ersten Konformitätsbewertung vorab bestimmt hat;

d) die in Artikel 14 genannten Maßnahmen zur Gewährleistung der menschlichen Aufsicht, einschließlich der technischen Maßnahmen, die getroffen wurden, um den Nutzern die Interpretation der Ergebnisse von KI-Systemen zu erleichtern;

e) die erforderlichen Rechen- und Hardware-Ressourcen, die erwartete Lebensdauer des Hochrisiko-KI-Systems und alle erforderlichen Wartungs- und Pflegemaßnahmen sowie deren Häufigkeit zur Gewährleistung des ordnungsgemäßen Funktionierens dieses KI-Systems, auch in Bezug auf Software-Updates;

f) eine Beschreibung des in das KI-System integrierten Mechanismus, der es den Nutzern gegebenenfalls ermöglicht, die Protokolle ordnungsgemäß zu erfassen, zu speichern und auszuwerten.

Artikel 14

Menschliche Aufsicht

(1) Hochrisiko-KI-Systeme werden so konzipiert und entwickelt, dass sie während der Dauer der Verwendung des KI-Systems – auch mit geeigneten Werkzeugen einer Mensch-Maschine-Schnittstelle – von natürlichen Personen wirksam beaufsichtigt werden können.

(2) Die menschliche Aufsicht dient der Verhinderung oder Minimierung der Risiken für die Gesundheit, die Sicherheit oder die Grundrechte, die entstehen können, wenn ein Hochrisiko-KI-System bestimmungsgemäß oder unter im Rahmen einer vernünftigerweise vorhersehbaren Fehlanwendung verwendet wird, insbesondere wenn solche Risiken trotz der Einhaltung anderer Anforderungen dieses Kapitels fortbestehen.

(3) Die menschliche Aufsicht wird durch eine oder alle der folgenden Arten von Vorkehrungen gewährleistet:
   a) Vorkehrungen, die vor dem Inverkehrbringen oder der Inbetriebnahme vom Anbieter bestimmt und, sofern technisch machbar, in das Hochrisiko-KI-System eingebaut werden;
   b) Vorkehrungen, die vor dem Inverkehrbringen oder der Inbetriebnahme des Hochrisiko-KI-Systems vom Anbieter bestimmt werden und dazu geeignet sind, vom Nutzer umgesetzt zu werden.

(4) Für die Zwecke der Umsetzung der Absätze 1 bis 3 wird das Hochrisiko-KI-System dem Nutzer so zur Verfügung gestellt, dass die natürlichen Personen, denen die menschliche Aufsicht übertragen wurde, je nach den Umständen und sofern verhältnismäßig in der Lage sind,
   a) die Fähigkeiten und Grenzen des Hochrisiko-KI-Systems zu verstehen und seinen Betrieb ordnungsgemäß zu überwachen;
   b) sich einer möglichen Neigung zu einem automatischen oder übermäßigen Vertrauen in das von einem Hochrisiko-KI-System hervorgebrachte Ergebnis („Automatisierungsbias“) bewusst zu bleiben;
   c) die Ergebnisse des Hochrisiko-KI-Systems richtig zu interpretieren, wobei beispielsweise die vorhandenen Interpretationswerkzeuge und -methoden zu berücksichtigen sind;
   d) in einer bestimmten Situation zu beschließen, das Hochrisiko-KI-System nicht zu verwenden oder das Ergebnis des Hochrisiko-KI-Systems außer Acht zu lassen, außer Kraft zu setzen oder rückgängig zu machen;

e) in den Betrieb des Hochrisiko-KI-Systems einzugreifen oder den Systembetrieb mit einer „Stopptaste“ oder einem ähnlichen Verfahren zu unterbrechen.

(5) Bei den in Anhang III Nummer 1 Buchstabe a genannten Hochrisiko-KI-Systemen müssen die in Absatz 3 genannten Vorkehrungen so gestaltet sein, dass außerdem der Nutzer keine Maßnahmen oder Entscheidungen allein aufgrund des vom System hervorgebrachten Identifizierungsergebnisses trifft, solange dies nicht von mindestens zwei natürlichen Personen getrennt überprüft und bestätigt wurde. Die Anforderung einer getrennten Überprüfung durch mindestens zwei natürliche Personen gilt nicht für Hochrisiko-KI-Systeme, die für Zwecke in den Bereichen Strafverfolgung, Migration, Grenzkontrolle oder Asyl verwendet werden, wenn die Anwendung dieser Anforderung nach Unionsrecht oder nationalem Recht unverhältnismäßig ist.

[...]

KAPITEL 3

PFLICHTEN DER ANBIETER UND NUTZER VON HOCHRISIKO-KI-SYSTEMEN UND ANDERER BETEILIGTER

Artikel 16

Pflichten der Anbieter von Hochrisiko-KI-Systemen

Anbieter von Hochrisiko-KI-Systemen müssen

a) sicherstellen, dass ihre Hochrisiko-KI-Systeme die Anforderungen in Kapitel 2 dieses Titels erfüllen;

aa) ihren Namen, ihren eingetragenen Handelsnamen oder ihre eingetragene Handelsmarke und ihre Kontaktanschrift auf dem Hochrisiko-KI-System selbst oder, wenn dies nicht möglich ist, auf der Verpackung oder in der beigefügten Dokumentation angeben;

b) über ein Qualitätsmanagementsystem verfügen, das dem Artikel 17 entspricht;

c) die in Artikel 18 genannte Dokumentation aufbewahren;

d) die von ihren Hochrisiko-KI-Systemen in Übereinstimmung mit Artikel 20 automatisch erzeugten Protokolle aufbewahren, wenn dies ihrer Kontrolle unterliegt;
e) sicherstellen, dass das Hochrisiko-KI-System dem betreffenden Konformitätsbewertungsverfahren nach Artikel 43 unterzogen wird, bevor es in Verkehr gebracht oder in Betrieb genommen wird;
f) den in Artikel 51 Absatz 1 genannten Registrierungspflichten nachkommen;
g) die erforderlichen Korrekturmaßnahmen gemäß Artikel 21 ergreifen, wenn das Hochrisiko-KI-System die Anforderungen in Kapitel 2 dieses Titels nicht erfüllt;
h) die betreffende zuständige nationale Behörde der Mitgliedstaaten, in denen sie das System bereitgestellt oder in Betrieb genommen haben, und gegebenenfalls die notifizierte Stelle über die Nichtkonformität und bereits ergriffene Korrekturmaßnahmen informieren;
i) die CE-Kennzeichnung an ihren Hochrisiko-KI-Systemen anbringen, um die Konformität mit dieser Verordnung gemäß Artikel 49 anzuzeigen;
j) auf Anfrage einer zuständigen nationalen Behörde nachweisen, dass das Hochrisiko-KI-System die Anforderungen in Kapitel 2 dieses Titels erfüllt

[...]

Artikel 19

Konformitätsbewertung

(1) Die Anbieter von Hochrisiko-KI-Systemen stellen sicher, dass ihre Systeme vor dem Inverkehrbringen oder der Inbetriebnahme dem betreffenden Konformitätsbewertungsverfahren gemäß Artikel 43 unterzogen werden. Wurde infolge dieser Konformitätsbewertung nachgewiesen, dass die KI-Systeme die Anforderungen in Kapitel 2 dieses Titels erfüllen, erstellen die Anbieter eine EU-Konformitätserklärung gemäß Artikel 48 und bringen die CE-Konformitätskennzeichnung gemäß Artikel 49 an.
(2) [gestrichen]

[...]

Artikel 23

Zusammenarbeit mit den zuständigen Behörden

Anbieter von Hochrisiko-KI-Systemen übermitteln einer zuständigen nationalen Behörde auf deren Anfrage alle Informationen und Unterlagen, die erforderlich sind, um die Konformität des Hochrisiko-KI-Systems mit den Anforderungen in Kapitel 2 dieses Titels nachzuweisen, in einer Sprache, die für die Behörde des betreffenden Mitgliedstaats leicht verständlich ist. Auf begründete Anfrage einer zuständigen nationalen Behörde gewähren die Anbieter dieser Behörde auch Zugang zu den von ihrem Hochrisiko-KI-System gemäß Artikel 12 Absatz 1 automatisch erzeugten Protokollen, soweit diese Protokolle aufgrund einer vertraglichen Vereinbarung mit dem Nutzer oder auf gesetzlicher Grundlage ihrer Kontrolle unterliegen.

[...]

Artikel 29

Pflichten der Nutzer von Hochrisiko-KI-Systemen

(1) Die Nutzer von Hochrisiko-KI-Systemen verwenden solche Systeme entsprechend der den Systemen beigefügten Gebrauchsanweisungen und gemäß den Absätzen 2 und 5 des vorliegenden Artikels.

(1a) Die Nutzer übertragen natürlichen Personen, die über die erforderliche Kompetenz, Ausbildung und Befugnis verfügen, die menschliche Aufsicht.

(2) Die Pflichten nach Absatz 1 und 1a lassen sonstige Pflichten der Nutzer nach Unionsrecht oder nationalem Recht sowie das Ermessen der Nutzer bei der Organisation ihrer eigenen Ressourcen und Tätigkeiten zur Wahrnehmung der vom Anbieter angegebenen Maßnahmen der menschlichen Aufsicht unberührt.

(3) Unbeschadet des Absatzes 1 und soweit die Eingabedaten seiner Kontrolle unterliegen, sorgen die Nutzer dafür, dass die Eingabedaten der Zweckbestimmung des Hochrisiko-KI-Systems entsprechen.

(4) Die Nutzer richten eine menschliche Aufsicht ein und überwachen den Betrieb des Hochrisiko-KI-Systems anhand der Gebrauchsanweisungen. Haben sie Grund zu der Annahme, dass die Verwendung gemäß den Ge-

brauchsanweisungen dazu führen kann, dass das Hochrisiko-KI-System ein Risiko im Sinne des Artikels 65 Absatz 1 birgt, so informieren sie den Anbieter oder Händler und setzen die Verwendung des Systems aus. Sie informieren den Anbieter oder Händler auch, wenn sie einen schwerwiegenden Vorfall festgestellt haben, und unterbrechen die Verwendung des KI-Systems. Kann der Nutzer den Anbieter nicht erreichen, so gilt Artikel 62 entsprechend. Diese Pflicht gilt nicht für sensible operative Daten von Nutzern von KI-Systemen, die Strafverfolgungsbehörden sind.

Bei Nutzern, die Finanzinstitute sind und gemäß den Rechtsvorschriften der Union im Bereich der Finanzdienstleistungen Anforderungen in Bezug auf ihre Regelungen oder Verfahren der internen Unternehmensführung unterliegen, gilt die in Unterabsatz 1 aufgeführte Überwachungspflicht als erfüllt, wenn die Vorschriften über Regelungen, Verfahren oder Mechanismen der internen Unternehmensführung gemäß den einschlägigen Rechtsvorschriften der Union im Bereich der Finanzdienstleistungen eingehalten werden.

(5) Nutzer von Hochrisiko-KI-Systemen bewahren die von ihrem Hochrisiko-KI-System gemäß Artikel 12 Absatz 1 automatisch erzeugten Protokolle auf, soweit diese Protokolle ihrer Kontrolle unterliegen. Sofern im geltenden Unionsrecht oder im nationalen Recht, insbesondere im Unionsrecht zum Schutz personenbezogener Daten, nichts anderes vorgesehen ist, bewahren sie sie mindestens sechs Monate lang auf.

Nutzer, die Finanzinstitute sind und gemäß den Rechtsvorschriften der Union im Bereich der Finanzdienstleistungen Anforderungen in Bezug auf ihre Regelungen oder Verfahren der internen Unternehmensführung unterliegen, bewahren die Protokolle als Teil der gemäß den einschlägigen Rechtsvorschriften der Union im Bereich der Finanzdienstleistungen aufzubewahrenden Dokumentation auf.

(5a) Nutzer von Hochrisiko-KI-Systemen, die Behörden, Einrichtungen oder sonstige Stellen sind, mit Ausnahme von Strafverfolgungs-, Grenzschutz-, Einwanderungs- oder Asylbehörden, erfüllen die in Artikel 51 genannten Registrierungspflichten. Stellen sie fest, dass das System, dessen Verwendung sie planen, nicht in der in Artikel 60 genannten EU-Datenbank registriert wurde, sehen sie von der Verwendung dieses Systems ab und unterrichten den Anbieter oder den Händler.

(6) Die Nutzer von Hochrisiko-KI-Systemen verwenden die gemäß Artikel 13 bereitgestellten Informationen, um gegebenenfalls ihrer Verpflichtung zur Durchführung einer Datenschutz-Folgenabschätzung gemäß Artikel 35 der Verordnung (EU) 2016/679 oder Artikel 27 der Richtlinie (EU) 2016/680 nachzukommen.

(6a) Die Nutzer arbeiten mit den zuständigen nationalen Behörden bei allen Maßnahmen zusammen, die diese Behörden im Zusammenhang mit einem von dem Nutzer verwendeten KI-System ergreifen.

KAPITEL 4

NOTIFIZIERENDE BEHÖRDEN UND NOTIFIZIERTE STELLEN

Artikel 30

Notifizierende Behörden

(1) Jeder Mitgliedstaat sorgt für die Benennung oder Schaffung mindestens einer notifizierenden Behörde, die für die Einrichtung und Durchführung der erforderlichen Verfahren zur Bewertung, Benennung und Notifizierung von Konformitätsbewertungsstellen und für deren Überwachung zuständig ist.

(2) Die Mitgliedstaaten können entscheiden, dass die Bewertung und Überwachung nach Absatz 1 von einer nationalen Akkreditierungsstelle im Sinne von und im Einklang mit der Verordnung (EG) Nr. 765/2008 erfolgt.

(3) Notifizierende Behörden werden so eingerichtet, strukturiert und in ihren Arbeitsabläufen organisiert, dass jegliche Interessenkonflikte mit Konformitätsbewertungsstellen vermieden werden und die Objektivität und die Unparteilichkeit ihrer Tätigkeiten gewährleistet sind.

(4) Notifizierende Behörden werden so strukturiert, dass Entscheidungen über die Notifizierung von Konformitätsbewertungsstellen von kompetenten Personen getroffen werden, die nicht mit den Personen identisch sind, die die Bewertung dieser Stellen durchgeführt haben.

(5) Notifizierende Behörden dürfen weder Tätigkeiten, die Konformitätsbewertungsstellen durchführen, noch Beratungsleistungen auf einer gewerblichen oder wettbewerblichen Basis anbieten oder erbringen.

(6) Notifizierende Behörden gewährleisten im Einklang mit Artikel 70 die Vertraulichkeit der von ihnen erlangten Informationen.
(7) Notifizierende Behörden verfügen über eine angemessene Anzahl kompetenter Mitarbeiter, sodass sie ihre Aufgaben ordnungsgemäß wahrnehmen können.
(8) [gestrichen]

Artikel 31

Antrag einer Konformitätsbewertungsstelle auf Notifizierung

(1) Konformitätsbewertungsstellen beantragen ihre Notifizierung bei der notifizierenden Behörde des Mitgliedstaats, in dem sie ansässig sind.
(2) Dem Antrag auf Notifizierung legen sie eine Beschreibung der Konformitätsbewertungstätigkeiten, des bzw. der Konformitätsbewertungsmoduls bzw. -module und der KI-Systeme, für die diese Konformitätsbewertungsstelle Kompetenz beansprucht, sowie, falls vorhanden, eine Akkreditierungsurkunde bei, die von einer nationalen Akkreditierungsstelle ausgestellt wurde und in der bescheinigt wird, dass die Konformitätsbewertungsstelle die Anforderungen des Artikels 33 erfüllt. Sonstige gültige Dokumente in Bezug auf bestehende Benennungen der antragstellenden notifizierten Stelle im Rahmen anderer Harmonisierungsrechtsvorschriften der Union sind ebenfalls beizufügen.
(3) Kann die Konformitätsbewertungsstelle keine Akkreditierungsurkunde vorweisen, so legt sie der notifizierenden Behörde als Nachweis alle Unterlagen vor, die erforderlich sind, um zu überprüfen, festzustellen und regelmäßig zu überwachen, ob sie die Anforderungen des Artikels 33 erfüllt. Bei notifizierten Stellen, die im Rahmen anderer Harmonisierungsrechtsvorschriften der Union benannt wurden, können alle Unterlagen und Bescheinigungen im Zusammenhang mit solchen Benennungen zur Unterstützung ihres Benennungsverfahrens nach dieser Verordnung verwendet werden. Die notifizierte Stelle aktualisiert die in Absätzen 2 und 3 genannten Unterlagen immer dann, wenn sich relevante Änderungen ergeben, damit die für notifizierte Stellen zuständige Behörde überwachen und überprüfen kann, ob die in Artikel 33 genannten Anforderungen kontinuierlich eingehalten werden.

Artikel 32

Notifizierungsverfahren

(1) Die notifizierenden Behörden dürfen nur Konformitätsbewertungsstellen notifizieren, die die Anforderungen von Artikel 33 erfüllen.
(2) Die notifizierenden Behörden unterrichten die Kommission und die anderen Mitgliedstaaten mithilfe des elektronischen Notifizierungsinstruments, das von der Kommission entwickelt und verwaltet wird, über diese Stellen.
(3) Eine Notifizierung gemäß Absatz 2 enthält vollständige Angaben zu den Konformitätsbewertungstätigkeiten, dem betreffenden Konformitätsbewertungsmodul oder den betreffenden Konformitätsbewertungsmodulen und des betreffenden KI-Systems sowie die einschlägige Bestätigung der Kompetenz. Beruht eine Notifizierung nicht auf einer Akkreditierungsurkunde gemäß Artikel 31 Absatz 2, so legt die notifizierende Behörde der Kommission und den anderen Mitgliedstaaten die Unterlagen, die die Kompetenz der Konformitätsbewertungsstelle nachweisen, sowie die Vereinbarungen vor, die getroffen wurden, um sicherzustellen, dass die Stelle regelmäßig überwacht wird und weiter stets den Anforderungen nach Artikel 33 genügt.
(4) Die betreffende Konformitätsbewertungsstelle darf die Tätigkeiten einer notifizierten Stelle nur dann wahrnehmen, wenn weder die Kommission noch die anderen Mitgliedstaaten innerhalb von zwei Wochen nach einer Notifizierung durch eine notifizierende Behörde, wenn eine Akkreditierungsurkunde gemäß Artikel 31 Absatz 2 vorgelegt wird, oder innerhalb von zwei Monaten nach einer Notifizierung durch eine notifizierende Behörde, wenn als Nachweis Unterlagen gemäß Artikel 31 Absatz 3 vorgelegt werden, Einwände erhoben haben.
(5) [gestrichen]

Artikel 33

Anforderungen an notifizierte Stellen

(1) Eine notifizierte Stelle muss nach nationalem Recht gegründet und mit Rechtspersönlichkeit ausgestattet sein.

(2) Die notifizierten Stellen müssen die Anforderungen an die Organisation, das Qualitätsmanagement, die Ressourcenausstattung und die Verfahren erfüllen, die zur Wahrnehmung ihrer Aufgaben erforderlich sind.
(3) Die Organisationsstruktur, die Zuweisung der Zuständigkeiten, die Berichtslinien und die Funktionsweise der notifizierten Stellen sind so gestaltet, dass das Vertrauen in die Leistung der notifizierten Stelle und in die Ergebnisse der von ihr durchgeführten Konformitätsbewertungstätigkeiten gewährleisten.
(4) Die notifizierten Stellen sind von dem Anbieter eines Hochrisiko-KI-Systems, zu dem sie Konformitätsbewertungstätigkeiten durchführen, unabhängig. Außerdem sind die notifizierten Stellen von allen anderen Akteuren, die ein wirtschaftliches Interesse an dem bewerteten Hochrisiko-KI-System haben, und von allen Wettbewerbern des Anbieters unabhängig.
(5) Die notifizierten Stellen gewährleisten durch ihre Organisation und Arbeitsweise, dass bei der Ausübung ihrer Tätigkeit Unabhängigkeit, Objektivität und Unparteilichkeit gewahrt sind. Von den notifizierten Stellen werden eine Struktur und Verfahren dokumentiert und umgesetzt, die ihre Unparteilichkeit gewährleisten und sicherstellen, dass die Grundsätze der Unparteilichkeit in ihrer gesamten Organisation und von allen Mitarbeitern und bei allen Bewertungstätigkeiten gefördert und angewandt werden.
(6) Die notifizierten Stellen gewährleisten durch dokumentierte Verfahren, dass ihre Mitarbeiter, Ausschüsse, Zweigstellen, Unterauftragnehmer sowie alle zugeordneten Stellen oder Mitarbeiter externer Einrichtungen im Einklang mit Artikel 70 die Vertraulichkeit der Informationen, die bei der Durchführung der Konformitätsbewertungstätigkeiten in ihren Besitz gelangen, wahren, außer wenn die Offenlegung gesetzlich vorgeschrieben ist. Informationen, von denen Mitarbeiter der notifizierten Stellen bei der Durchführung ihrer Aufgaben gemäß dieser Verordnung Kenntnis erlangen, unterliegen der beruflichen Schweigepflicht, außer gegenüber den notifizierenden Behörden des Mitgliedstaats, in dem sie ihre Tätigkeiten ausüben.
(7) Die notifizierten Stellen verfügen über Verfahren zur Durchführung ihrer Tätigkeiten unter gebührender Berücksichtigung der Größe eines Unternehmens, der Branche, in der es tätig ist, seiner Struktur sowie der Komplexität des betreffenden KI-Systems.

(8) Die notifizierten Stellen schließen eine angemessene Haftpflichtversicherung für ihre Konformitätsbewertungstätigkeiten ab, es sei denn, diese Haftpflicht wird aufgrund nationalen Rechts von dem Mitgliedstaat, in dem sie sich befinden, gedeckt oder dieser Mitgliedstaat ist selbst unmittelbar für die Durchführung der Konformitätsbewertung zuständig.

(9) Die notifizierten Stellen sind in der Lage, die ihnen durch diese Verordnung zufallenden Aufgaben mit höchster beruflicher Integrität und der erforderlichen Fachkompetenz in dem betreffenden Bereich auszuführen, gleichgültig, ob diese Aufgaben von den notifizierten Stellen selbst oder in ihrem Auftrag und in ihrer Verantwortung erfüllt werden.

(10) Die notifizierten Stellen verfügen über ausreichende interne Kompetenzen, um die von externen Stellen in ihrem Namen wahrgenommen Aufgaben wirksam beurteilen zu können. Die notifizierten Stellen verfügen ständig über ausreichendes administratives, technisches, juristisches und wissenschaftliches Personal, das die entsprechenden Erfahrungen und Kenntnisse in Bezug auf einschlägige KI-Technik, Daten und Datenverarbeitung sowie die Anforderungen in Kapitel 2 dieses Titels besitzt.

(11) Die notifizierten Stellen wirken an den in Artikel 38 genannten Koordinierungstätigkeiten mit. Sie wirken außerdem unmittelbar oder mittelbar an der Arbeit der europäischen Normungsorganisationen mit oder stellen sicher, dass sie stets über den Stand der einschlägigen Normen unterrichtet sind.

(12) [gestrichen]

[...]

Artikel 34

Zweigstellen notifizierter Stellen und Vergabe von Unteraufträgen durch notifizierte Stellen

(1) Vergibt die notifizierte Stelle bestimmte mit der Konformitätsbewertung verbundene Aufgaben an Unterauftragnehmer oder überträgt sie diese einer Zweigstelle, so stellt sie sicher, dass der Unterauftragnehmer oder die Zweigstelle die Anforderungen des Artikels 33 erfüllt, und setzt die notifizierende Behörde davon in Kenntnis.

(2) Die notifizierten Stellen tragen die volle Verantwortung für die Arbeiten, die von Unterauftragnehmern oder Zweigstellen ausgeführt werden, unabhängig davon, wo diese niedergelassen sind.
(3) Arbeiten dürfen nur mit Zustimmung des Anbieters an einen Unterauftragnehmer vergeben oder einer Zweigstelle übertragen werden.
(4) Die einschlägigen Unterlagen über die Bewertung der Qualifikation des Unterauftragnehmers oder der Zweigstelle und die von ihnen gemäß dieser Verordnung ausgeführten Arbeiten werden für einen Zeitraum von fünf Jahren ab dem Datum der Beendigung der Vergabe von Unteraufträgen für die notifizierende Behörde bereitgehalten.

[...]

Artikel 43

Konformitätsbewertung

(1) Hat ein Anbieter zum Nachweis, dass sein in Anhang III Nummer 1 aufgeführtes Hochrisiko-KI-System die Anforderungen in Kapitel 2 dieses Titels erfüllt, harmonisierte Normen gemäß Artikel 40 oder gegebenenfalls gemeinsame Spezifikationen gemäß Artikel 41 angewandt, so entscheidet er sich für eines der folgenden Verfahren:
   a) das Konformitätsbewertungsverfahren auf der Grundlage einer internen Kontrolle gemäß Anhang VI oder
   b) das Konformitätsbewertungsverfahren auf der Grundlage der Bewertung des Qualitätsmanagementsystems und der Bewertung der technischen Dokumentation unter Beteiligung einer notifizierten Stelle gemäß Anhang VII.

   Hat ein Anbieter zum Nachweis, dass sein Hochrisiko-KI-System die Anforderungen in Kapitel 2 dieses Titels erfüllt, die harmonisierten Normen gemäß Artikel 40 nicht oder nur teilweise angewandt oder gibt es solche harmonisierten Normen nicht und liegen keine gemeinsamen Spezifikationen gemäß Artikel 41 vor, so befolgt er das Konformitätsbewertungsverfahren gemäß Anhang VII.

   Für die Zwecke des Konformitätsbewertungsverfahrens gemäß Anhang VII kann der Anbieter eine der notifizierten Stellen auswählen. Soll das System jedoch von Strafverfolgungs-, Einwanderungs- oder Asylbehörden

oder von Organen, Einrichtungen oder sonstigen Stellen der EU in Betrieb genommen werden, so übernimmt die in Artikel 63 Absatz 5 oder 6 genannte Marktüberwachungsbehörde die Funktion der notifizierten Stelle.

(2) Bei den in Anhang III Nummern 2 bis 8 aufgeführten Hochrisiko-KI-Systemen und bei den in Titel 1a genannten KI-Systemen mit allgemeinem Verwendungszweck befolgen die Anbieter das Konformitätsbewertungsverfahren auf der Grundlage einer internen Kontrolle gemäß Anhang VI, das keine Beteiligung einer notifizierten Stelle vorsieht.

(3) Bei den Hochrisiko-KI-Systemen, die unter die in Anhang II Abschnitt A aufgeführten Rechtsakte fallen, befolgt der Anbieter die einschlägigen Konformitätsbewertungsverfahren, die nach diesen Rechtsakten erforderlich sind. Die Anforderungen in Kapitel 2 dieses Titels gelten für diese Hochrisiko-KI-Systeme und werden in diese Bewertung einbezogen. Anhang VII Nummern 4.3, 4.4, 4.5 und Nummer 4.6 Absatz 5 finden ebenfalls Anwendung.

Für die Zwecke dieser Bewertung sind die notifizierten Stellen, die gemäß diesen Rechtsakten benannt wurden, auch berechtigt, die Konformität der Hochrisiko-KI-Systeme mit den Anforderungen in Kapitel 2 dieses Titels zu kontrollieren, sofern im Rahmen des gemäß diesen Rechtsakten durchgeführten Notifizierungsverfahrens geprüft wurde, dass diese notifizierten Stellen die in Artikel 33 Absätze 4, 9 und 10 festgelegten Anforderungen erfüllen.

Wenn die in Anhang II Abschnitt A aufgeführten Rechtsakte es dem Hersteller des Produkts ermöglichen, auf eine Konformitätsbewertung durch Dritte zu verzichten, sofern dieser Hersteller alle harmonisierten Normen, die alle einschlägigen Anforderungen abdecken, angewandt hat, so darf dieser Hersteller nur dann von dieser Möglichkeit Gebrauch machen, wenn er auch harmonisierte Normen oder gegebenenfalls gemeinsame Spezifikationen gemäß Artikel 41, die die Anforderungen in Kapitel 2 dieses Titels abdecken, angewandt hat.

(4) [gestrichen]

(5) Der Kommission wird die Befugnis übertragen, gemäß Artikel 73 delegierte Rechtsakte zu erlassen, um die Anhänge VI und VII angesichts des technischen Fortschritts zu aktualisieren.

(6) Der Kommission wird die Befugnis übertragen, delegierte Rechtsakte zur Änderung der Absätze 1 und 2 zu erlassen, um die in Anhang III Nummern

2 bis 8 genannten Hochrisiko-KI-Systeme dem Konformitätsbewertungsverfahren gemäß Anhang VII oder Teilen davon zu unterwerfen. Die Kommission erlässt solche delegierten Rechtsakte unter Berücksichtigung der Wirksamkeit des Konformitätsbewertungsverfahrens auf der Grundlage einer internen Kontrolle gemäß Anhang VI hinsichtlich der Vermeidung oder Minimierung der von solchen Systemen ausgehenden Risiken für die Gesundheit und Sicherheit und den Schutz der Grundrechte sowie hinsichtlich der Verfügbarkeit angemessener Kapazitäten und Ressourcen in den notifizierten Stellen.

[...]

TITEL IV

TRANSPARENZPFLICHTEN FÜR ANBIETER UND NUTZER BESTIMMTER KI-SYSTEME

Artikel 52

Transparenzpflichten für Anbieter und Nutzer bestimmter KI-Systeme

(1) Die Anbieter stellen sicher, dass KI-Systeme, die für die Interaktion mit natürlichen Personen bestimmt sind, so konzipiert und entwickelt werden, dass natürlichen Personen mitgeteilt wird, dass sie es mit einem KI-System zu tun haben, es sei denn, dies ist aus Sicht einer normal informierten, angemessen aufmerksamen, verständigen natürlichen Person aufgrund der Umstände und des Kontexts der Nutzung offensichtlich. Diese Vorgabe gilt nicht für gesetzlich zur Aufdeckung, Verhütung, Ermittlung und Verfolgung von Straftaten zugelassene KI-Systeme, wenn geeignete Schutzvorkehrungen für die Rechte und Freiheiten Dritter bestehen, es sei denn, diese Systeme stehen der Öffentlichkeit zur Anzeige einer Straftat zur Verfügung.

(2) Die Nutzer eines Systems zur biometrischen Kategorisierung informieren die davon betroffenen natürlichen Personen über den Betrieb des Systems. Diese Vorgabe gilt nicht für gesetzlich zur Aufdeckung, Verhütung, Ermittlung und Verfolgung von Straftaten zugelassene KI-Systeme, die zur

biometrischen Kategorisierung verwendet werden, wenn geeignete Schutzvorkehrungen für die Rechte und Freiheiten Dritter bestehen.

(2a) Die Nutzer eines Emotionserkennungssystems informieren die davon betroffenen natürlichen Personen über den Betrieb des Systems. Diese Vorgabe gilt nicht für gesetzlich zur Aufdeckung, Verhütung, Ermittlung und Verfolgung von Straftaten zugelassene KI-Systeme, die als Emotionserkennungssysteme eingesetzt werden, wenn geeignete Schutzvorkehrungen für die Rechte und Freiheiten Dritter bestehen.

(3) Nutzer eines KI-Systems, das Bild-, Ton- oder Videoinhalte erzeugt oder manipuliert, die wirklichen Personen, Gegenständen, Orten oder anderen Einrichtungen oder Ereignissen merklich ähneln und einer Person fälschlicherweise als echt oder wahrhaftig erscheinen würden („Deepfake"), müssen offenlegen, dass die Inhalte künstlich erzeugt oder manipuliert wurden. Unterabsatz 1 gilt jedoch nicht, wenn die Verwendung zur Aufdeckung, Verhütung, Ermittlung und Verfolgung von Straftaten gesetzlich zugelassen oder der Inhalt Teil eines offensichtlich kreativen, satirischen, künstlerischen oder fiktionalen Werks oder Programms ist und geeignete Schutzvorkehrungen für die Rechte und Freiheiten Dritter bestehen.

(3a) Die in den Absätzen 1 bis 3 genannten Informationen werden den natürlichen Personen spätestens zum Zeitpunkt der ersten Interaktion oder Aussetzung in klarer und eindeutiger Weise bereitgestellt.

(4) Die Absätze 1, 2, 2a, 3 und 3a lassen die in Titel III dieser Verordnung festgelegten Anforderungen und Pflichten sowie andere im Unionsrecht oder im einzelstaatlichen Recht festlegte Transparenzpflichten für Nutzer von KI-Systemen unberührt.

[...]

KAPITEL 3

DURCHSETZUNG

Artikel 63

Marktüberwachung und Kontrolle von KI-Systemen auf dem Unionsmarkt

(1) Die Verordnung (EU) 2019/1020 gilt für KI-Systeme, die unter diese Verordnung fallen. Für die Zwecke einer wirksamen Durchsetzung dieser Verordnung gilt jedoch Folgendes:
   a) Jede Bezugnahme auf einen Wirtschaftsakteur nach der Verordnung (EU) 2019/1020 gilt auch als Bezugnahme auf alle Akteure, die in Artikel 2 dieser Verordnung genannt werden.
   b) Jede Bezugnahme auf ein Produkt nach der Verordnung (EU) 2019/1020 gilt auch als Bezugnahme auf alle KI-Systeme, die unter diese Verordnung fallen.

(2) Die Marktüberwachungsbehörden erstatten der Kommission im Rahmen ihrer Meldepflichten gemäß Artikel 34 Absatz 4 der Verordnung (EU) 2019/1020 über die Ergebnisse ihrer jeweiligen Marktüberwachungstätigkeiten gemäß dieser Verordnung Bericht.

(3) Bei Hochrisiko-KI-Systemen und damit in Zusammenhang stehenden Produkten, auf die die in Anhang II Abschnitt A aufgeführten Rechtsakte Anwendung finden, gilt als Marktüberwachungsbehörde für die Zwecke dieser Verordnung die in jenen Rechtsakten für die Marktüberwachung benannte Behörde oder – in begründeten Fällen und wenn für Abstimmung gesorgt ist – eine andere von dem Mitgliedstaat benannte einschlägige Behörde.

   Die Verfahren gemäß den Artikeln 65, 66, 67 und 68 dieser Verordnung gelten nicht für KI-Systeme für Produkte, die unter die in Anhang II Abschnitt A aufgeführten Rechtsakte fallen, wenn in diesen Rechtsakten bereits Verfahren mit demselben Ziel vorgesehen sind. In diesem Fall kommen die sektorspezifischen Verfahren zur Anwendung.

(4) Bei Hochrisiko-KI-Systemen, die von auf der Grundlage des Finanzdienstleistungsrechts der Union regulierten Finanzinstituten in Verkehr gebracht, in Betrieb genommen oder verwendet werden, gilt die in jenen

Rechtsvorschriften für die Finanzaufsicht über diese Institute benannte nationale Behörde als Marktüberwachungsbehörde für die Zwecke dieser Verordnung, sofern das Inverkehrbringen, die Inbetriebnahme oder die Verwendung des KI-Systems mit der Erbringung dieser Finanzdienstleistungen in direktem Zusammenhang steht.

Abweichend vom vorangehenden Unterabsatz kann der Mitgliedstaat – in begründeten Fällen und wenn für Abstimmung gesorgt ist – eine andere einschlägige Behörde als Marktüberwachungsbehörde für die Zwecke dieser Verordnung benennen.

Nationale Marktüberwachungsbehörden, die auf der Grundlage der Richtlinie 2013/36/EU regulierte Kreditinstitute, die an dem mit der Verordnung (EU) Nr. 1042/2013 des Rates eingerichteten einheitlichen Aufsichtsmechanismus teilnehmen, beaufsichtigen, sollten der Europäischen Zentralbank unverzüglich alle im Zuge ihrer Marktüberwachungstätigkeiten ermittelten Informationen übermitteln, die für die in der genannten Verordnung festgelegten Aufsichtsaufgaben der Europäischen Zentralbank von Belang sein könnten.

(5) Für die in Absatz 1 Buchstabe a genannten Hochrisiko-KI-Systeme, sofern diese Systeme für Strafverfolgungszwecke nach Anhang III Nummern 6, 7 und 8 verwendet werden, benennen die Mitgliedstaaten entweder die nationalen Behörden, die die Tätigkeiten der Strafverfolgungs-, Grenzschutz-, Einwanderungs-, Asyl- oder Justizbehörden beaufsichtigen, oder die für den Datenschutz nach der Richtlinie (EU) 2016/680 oder der Verordnung (EU) 2016/679 zuständigen Aufsichtsbehörden als Marktüberwachungsbehörden für die Zwecke dieser Verordnung. Marktüberwachungstätigkeiten dürfen in keiner Weise Auswirkungen auf die Unabhängigkeit von Justizbehörden haben oder deren Handlungen im Rahmen ihrer justiziellen Tätigkeit anderweitig beeinflussen.

(6) Soweit Organe, Einrichtungen und sonstige Stellen der Union in den Anwendungsbereich dieser Verordnung fallen, übernimmt der Europäische Datenschutzbeauftragte die Funktion der für sie zuständigen Marktüberwachungsbehörde.

(7) Die Mitgliedstaaten erleichtern die Koordinierung zwischen den auf der Grundlage dieser Verordnung benannten Marktüberwachungsbehörden und anderen einschlägigen nationalen Behörden oder Stellen, die die Anwendung der in Anhang II aufgeführten Harmonisierungsrechtsvorschrif-

ten der Union oder sonstigen Unionsrechts überwachen, das für die in Anhang III aufgeführten Hochrisiko-KI-Systeme relevant sein könnte.

(8) Der Anbieter gewährt den Marktüberwachungsbehörden unbeschadet der Befugnisübertragung gemäß der Verordnung (EU) 2019/1020, sofern dies relevant ist und beschränkt auf das zur Wahrnehmung der Aufgaben dieser Behörden erforderliche Maß, uneingeschränkten Zugang zur Dokumentation sowie zu den für die Entwicklung des Hochrisiko-KI-Systems verwendeten Trainings-, Validierungs- und Testdatensätzen, einschließlich, sofern dies relevant ist und im Rahmen der Sicherheitsmaßnahmen, über die Anwendungsprogrammierschnittstellen (API) oder andere einschlägige technische Mittel und Tools, die den Fernzugriff ermöglichen.

(9) Zum Quellcode des Hochrisiko-KI-Systems erhalten Marktüberwachungsbehörden auf begründete Anfrage und nur dann Zugang, wenn die folgenden kumulativen Bedingungen erfüllt sind:
   a) Der Zugang zum Quellcode ist zur Bewertung der Konformität eines Hochrisiko-KI-Systems mit den in Titel III Kapitel 2 festgelegten Anforderungen notwendig, und
   b) die Test-/Prüfverfahren und Überprüfungen aufgrund der vom Anbieter bereitgestellten Daten und Dokumentation wurden ausgeschöpft oder haben sich als unzureichend erwiesen.

(10) Jegliche Informationen und Dokumentation, in deren Besitz die Marktüberwachungsbehörden gelangt, werden im Einklang mit den in Artikel 70 festgelegten Vertraulichkeitspflichten behandelt.

(11) Natürliche oder juristische Personen, die Grund zu der Annahme haben, dass gegen die Bestimmungen dieser Verordnung verstoßen wurde, können bei der betreffenden Marktüberwachungsbehörde Beschwerde einlegen. Gemäß Artikel 11 Absatz 3 Buchstabe e und Absatz 7 Buchstabe a der Verordnung (EU) 2019/1020 werden Beschwerden für die Zwecke der Durchführung von Marktüberwachungstätigkeiten berücksichtigt und nach den einschlägigen, von den Marktüberwachungsbehörden dafür eingerichteten Verfahren behandelt.

[...]

Artikel 64

Befugnisse der für den Schutz der Grundrechte zuständigen Behörden

(1) [gestrichen]
(2) [gestrichen]
(3) Nationale Behörden oder öffentliche Stellen, die die Einhaltung des Unionsrechts zum Schutz der Grundrechte, einschließlich des Rechts auf Nichtdiskriminierung, in Bezug auf die Verwendung der in Anhang III aufgeführten Hochrisiko-KI-Systeme überwachen oder durchsetzen, sind befugt, alle auf der Grundlage dieser Verordnung erstellten oder geführten Unterlagen anzufordern und einzusehen, sofern der Zugang zu diesen Unterlagen für die Ausübung ihres Auftrags im Rahmen ihrer Befugnisse notwendig ist. Die jeweilige Behörde oder öffentliche Stelle unterrichtet die Marktüberwachungsbehörde des betreffenden Mitgliedstaats von jeder diesbezüglichen Anfrage.
(4) Bis drei Monate nach dem Inkrafttreten dieser Verordnung muss jeder Mitgliedstaat die in Absatz 3 genannten Behörden oder öffentlichen Stellen benannt haben und deren Liste veröffentlichen. Die Mitgliedstaaten übermitteln die Liste der Kommission und allen anderen Mitgliedstaaten und sorgen dafür, dass die Liste stets aktuell bleibt.
(5) Sollte die in Absatz 3 genannte Dokumentation nicht ausreichen, um feststellen zu können, ob ein Verstoß gegen das Unionsrecht zum Schutz der Grundrechte vorliegt, kann die in Absatz 3 genannte Behörde oder öffentliche Stelle bei der Marktüberwachungsbehörde einen begründeten Antrag auf Durchführung technischer Tests des Hochrisiko-KI-Systems stellen. Die Marktüberwachungsbehörde führt den Test unter enger Einbeziehung der beantragenden Behörde oder öffentlichen Stelle innerhalb eines angemessenen Zeitraums nach Eingang des Antrags durch.
(6) Alle Informationen und Unterlagen, in deren Besitz eine in Absatz 3 genannte nationale Behörde oder öffentliche Stelle auf der Grundlage dieses Artikels gelangt, werden im Einklang mit den in Artikel 70 festgelegten Vertraulichkeitspflichten behandelt.

[...]

TITEL X

VERTRAULICHKEIT UND SANKTIONEN

Artikel 70

Vertraulichkeit

(1) Die zuständigen nationalen Behörden, die notifizierten Stellen, die Kommission, der KI-Ausschuss und alle anderen natürlichen oder juristischen Personen, die an der Anwendung dieser Verordnung beteiligt sind, ergreifen im Einklang mit dem Unionsrecht oder dem nationalen Recht geeignete technische und organisatorische Maßnahmen, um die Vertraulichkeit der Informationen und Daten, in deren Besitz sie bei der Ausführung ihrer Aufgaben und Tätigkeiten gelangen, sicherzustellen, sodass insbesondere Folgendes geschützt ist:
   a) Rechte des geistigen Eigentums, vertrauliche Geschäftsinformationen oder Geschäftsgeheimnisse natürlicher oder juristischer Personen, auch Quellcodes, mit Ausnahme der in Artikel 5 der Richtlinie 2016/943 über den Schutz vertraulichen Know-hows und vertraulicher Geschäftsinformationen (Geschäftsgeheimnisse) vor rechtswidrigem Erwerb sowie rechtswidriger Nutzung und Offenlegung genannten Fälle;
   b) die wirksame Umsetzung dieser Verordnung, insbesondere für die Zwecke von Inspektionen, Untersuchungen oder Audits,
   c) öffentliche und nationale Sicherheitsinteressen;
   d) die Integrität von Straf- oder Verwaltungsverfahren.
   e) die Integrität von gemäß dem Unionsrecht oder dem nationalen Recht als Verschlusssachen eingestuften Informationen.

(2) Unbeschadet des Absatzes 1 darf der Austausch vertraulicher Informationen zwischen den zuständigen nationalen Behörden untereinander sowie zwischen den zuständigen nationalen Behörden und der Kommission nicht ohne vorherige Rücksprache mit der zuständigen nationalen Behörde und dem Nutzer, von der bzw. dem die Informationen stammen, offengelegt werden, sofern die Hochrisiko-KI-Systeme nach Anhang III Nummern 1, 6 und 7 von Strafverfolgungs-, Grenzschutz-, Einwanderungs- oder Asylbehörden verwendet werden und eine solche Offenlegung die öffentlichen und nationalen Sicherheitsinteressen gefährden könnte. Diese Pflicht zum

Austausch von Informationen erstreckt sich nicht auf sensible operative Daten zu den Tätigkeiten von Strafverfolgungs-, Grenzschutz-, Einwanderungs- oder Asylbehörden.

Handeln Strafverfolgungs-, Einwanderungs- oder Asylbehörden als Anbieter von Hochrisiko-KI-Systemen, wie sie in Anhang III Nummern 1, 6 und 7 aufgeführt sind, so verbleibt die technische Dokumentation nach Anhang IV in den Räumlichkeiten dieser Behörden. Diese Behörden müssen dafür sorgen, dass die in Artikel 63 Absätze 5 bzw. 6 genannten Marktüberwachungsbehörden auf Anfrage unverzüglich Zugang zu dieser Dokumentation oder eine Kopie davon erhalten. Zugang zu dieser Dokumentation oder zu einer Kopie davon darf nur das Personal der Marktüberwachungsbehörde erhalten, das über eine entsprechende Sicherheitsfreigabe verfügt.

(3) Die Absätze 1 und 2 dürfen sich weder auf die Rechte und Pflichten der Kommission, der Mitgliedstaaten, ihrer einschlägigen Behörden sowie der notifizierten Stellen in Bezug auf den Informationsaustausch und die Weitergabe von Warnungen, auch im Rahmen der grenzüberschreitenden Zusammenarbeit, noch auf die Pflichten der betreffenden Parteien auswirken, Informationen auf der Grundlage des Strafrechts der Mitgliedstaaten bereitzustellen.

[...]

ANHANG IV

TECHNISCHE DOKUMENTATION GEMÄß ARTIKEL 11 ABSATZ 1

Die in Artikel 11 Absatz 1 genannte technische Dokumentation muss mindestens die folgenden Informationen enthalten, soweit sie für das betreffende KI-System von Belang sind:

1. Allgemeine Beschreibung des KI-Systems, einschließlich
   a) Zweckbestimmung, das System entwickelnde Person(en), Datum und Version des Systems
   b) gegebenenfalls Interaktion oder Verwendung des KI-Systems mit Hardware oder Software, die nicht Teil des KI-Systems selbst sind
   c) Versionen der betreffenden Software oder Firmware und etwaige Anforderungen in Bezug auf die Aktualisierung der Versionen

d) Beschreibung aller Formen, in denen das KI-System in Verkehr gebracht oder in Betrieb genommen wird (z. B. in Hardware eingebettetes Softwarepaket, herunterladbar, API)
e) Beschreibung der Hardware, auf der das KI-System betrieben werden soll
f) falls das KI-System Bestandteil von Produkten ist: Fotografien oder Abbildungen, die äußere Merkmale, Kennzeichnungen und den inneren Aufbau dieser Produkte zeigen
g) Gebrauchsanweisungen für die Nutzer und gegebenenfalls Aufbau- oder Installationsanweisungen

2. Detaillierte Beschreibung der Bestandteile des KI-Systems und seines Entwicklungsprozesses, einschließlich
   a) Methoden und Schritte zur Entwicklung des KI-Systems, gegebenenfalls einschließlich des Einsatzes von Dritten bereitgestellter vortrainierter Systeme oder Werkzeuge, und wie diese vom Anbieter benutzt, integriert oder verändert wurden
   b) Entwurfsspezifikationen des Systems, insbesondere die allgemeine Logik des KI-Systems und der Algorithmen; wichtigste Entwurfsentscheidungen mit den Gründen und Annahmen, auch in Bezug auf Personen oder Personengruppen, auf die das System angewandt werden soll; hauptsächliche Klassifizierungsentscheidungen; was das System optimieren soll und welche Bedeutung den verschiedenen Parametern dabei zukommt; Beschreibung des erwarteten Ergebnisses des Systems; Entscheidungen über mögliche Kompromisse in Bezug auf die technischen Lösungen, mit denen die Anforderungen in Titel III Kapitel 2 erfüllt werden sollen
   c) Beschreibung der Systemarchitektur, aus der hervorgeht, wie Softwarekomponenten aufeinander aufbauen oder einander zuarbeiten und in die Gesamtverarbeitung integriert sind; zum Entwickeln, Trainieren, Testen und Validieren des KI-Systems verwendete Rechenressourcen
   d) gegebenenfalls Datenanforderungen in Form von Datenblättern, in denen die Trainingsmethoden und -techniken und die verwendeten Trainingsdatensätze beschrieben werden, einschließlich einer allgemeinen Beschreibung dieser Datensätze sowie Angaben zu deren Herkunft, Umfang und Hauptmerkmalen; Angaben zur Beschaffung und Aus-

wahl der Daten; Kennzeichnungsverfahren (z. B. für überwachtes Lernen), Datenbereinigungsmethoden (z. B. Erkennung von Ausreißern)

e) Bewertung der nach Artikel 14 erforderlichen Maßnahmen der menschlichen Aufsicht, mit einer Bewertung der technischen Maßnahmen, die erforderlich sind, um den Nutzern gemäß Artikel 13 Absatz 3 Buchstabe d die Interpretation der Ergebnisse von KI-Systemen zu erleichtern

f) gegebenenfalls detaillierte Beschreibung der vorab bestimmten Änderungen an dem KI-System und seiner Leistung mit allen einschlägigen Angaben zu den technischen Lösungen, mit denen sichergestellt wird, dass das KI-System die einschlägigen Anforderungen nach Titel III Kapitel 2 weiterhin dauerhaft erfüllt

g) verwendete Validierungs- und Testverfahren, mit Angaben zu den verwendeten Validierungs- und Testdaten und deren Hauptmerkmalen; Parameter, die zur Messung der Genauigkeit, Robustheit, Cybersicherheit und der Erfüllung anderer einschlägiger Anforderungen nach Titel III Kapitel 2 sowie potenziell diskriminierender Auswirkungen verwendet werden; Testprotokolle und alle von den verantwortlichen Personen datierten und unterzeichneten Testberichte, auch in Bezug auf die in Buchstabe f genannten vorab bestimmten Änderungen

3. Detaillierte Informationen über die Überwachung, Funktionsweise und Kontrolle des KI-Systems, insbesondere in Bezug auf: die Fähigkeiten und Leistungsgrenzen des Systems, einschließlich seines Genauigkeitsgrads bei bestimmten Personen oder Personengruppen, auf die es bestimmungsgemäß angewandt werden soll, sowie des in Bezug auf seine Zweckbestimmung insgesamt erwarteten Genauigkeitsgrads; angesichts der Zweckbestimmung des KI-Systems vorhersehbare unbeabsichtigte Ergebnisse und Risikoquellen für die Gesundheit und Sicherheit, die Grundrechte und eine etwaige Diskriminierung; die nach Artikel 14 erforderlichen Maßnahmen der menschlichen Aufsicht, einschließlich der technischen Maßnahmen, die getroffen wurden, um den Nutzern die Interpretation der Ergebnisse von KI-Systemen zu erleichtern; gegebenenfalls Spezifikationen zu Eingabedaten
4. Detaillierte Beschreibung des Risikomanagementsystems gemäß Artikel 9
5. Beschreibung einschlägiger Änderungen, die der Anbieter während des Lebenszyklus an dem System vorgenommen hat

6. Aufstellung der vollständig oder teilweise angewandten harmonisierten Normen, deren Fundstellen im Amtsblatt der Europäischen Union veröffentlicht worden sind; falls keine solchen harmonisierten Normen angewandt werden, eine detaillierte Beschreibung der Lösungen, mit denen die Anforderungen in Titel III Kapitel 2 erfüllt werden sollen, mit einer Aufstellung anderer einschlägiger Normen und technischer Spezifikationen
7. Kopie der EU-Konformitätserklärung
8. Detaillierte Beschreibung des Systems zur Bewertung der Leistung des KI-Systems in der Phase nach dem Inverkehrbringen gemäß Artikel 61, mit dem in Artikel 61 Absatz 3 genannten Plan für die Beobachtung nach dem Inverkehrbringen.

[...]

ANHANG VI

KONFORMITÄTSBEWERTUNGSVERFAHREN AUF DER GRUNDLAGE EINER INTERNEN KONTROLLE

1. Das Konformitätsbewertungsverfahren auf der Grundlage einer internen Kontrolle ist das Konformitätsbewertungsverfahren gemäß den Nummern 2 bis 4.
2. Der Anbieter überprüft, ob das bestehende Qualitätsmanagementsystem den Anforderungen des Artikels 17 entspricht.
3. Der Anbieter prüft die in der technischen Dokumentation enthaltenen Informationen, um zu beurteilen. ob das KI-System den einschlägigen grundlegenden Anforderungen in Titel III Kapitel 2 entspricht.
4. Der Anbieter überprüft ferner, ob der Entwurfs- und Entwicklungsprozess des KI-Systems und seine Beobachtung nach dem Inverkehrbringen gemäß Artikel 61 mit der technischen Dokumentation im Einklang stehen.

ANHANG VII

KONFORMITÄT AUF DER GRUNDLAGE DER BEWERTUNG DES QUALITÄTSMANAGEMENTSYSTEMS UND DER BEWERTUNG DER TECHNISCHEN DOKUMENTATION

1. Einleitung
   Das Konformitätsbewertungsverfahren auf der Grundlage der Bewertung des Qualitätsmanagementsystems und der Bewertung der technischen Dokumentation ist das Konformitätsbewertungsverfahren gemäß den Nummern 2 bis 5.
2. Überblick
   Das genehmigte Qualitätsmanagementsystem für die Konzeption, die Entwicklung und das Testen von KI-Systemen nach Artikel 17 wird gemäß Nummer 3 geprüft und unterliegt der Überwachung gemäß Nummer 5. Die technische Dokumentation des KI-Systems wird gemäß Nummer 4 geprüft.
3. Qualitätsmanagementsystem

3.1. Der Antrag des Anbieters muss Folgendes enthalten:
   a) den Namen und die Anschrift des Anbieters sowie, wenn der Antrag vom Bevollmächtigten eingereicht wird, auch dessen Namen und Anschrift,
   b) die Liste der unter dasselbe Qualitätsmanagementsystem fallenden KI-Systeme,
   c) die technische Dokumentation für jedes unter dasselbe Qualitätsmanagementsystem fallende KI-System,
   d) die Dokumentation über das Qualitätsmanagementsystem mit allen in Artikel 17 aufgeführten Aspekten,
   e) eine Beschreibung der bestehenden Verfahren, mit denen sichergestellt wird, dass das Qualitätsmanagementsystem geeignet und wirksam bleibt,
   f) eine schriftliche Erklärung, dass derselbe Antrag bei keiner anderen notifizierten Stelle eingereicht worden ist.

3.2. Das Qualitätssicherungssystem wird von der notifizierten Stelle bewertet, um festzustellen, ob es die in Artikel 17 genannten Anforderungen erfüllt. Die Entscheidung wird dem Anbieter oder dessen Bevollmächtigten mitgeteilt.

Die Mitteilung enthält die Ergebnisse der Bewertung des Qualitätsmanagementsystems und die begründete Bewertungsentscheidung.

3.3. Das genehmigte Qualitätsmanagementsystem wird vom Anbieter weiter angewandt und gepflegt, damit es stets sachgemäß und effizient funktioniert.

3.4. Der Anbieter unterrichtet die notifizierte Stelle über jede beabsichtigte Änderung des genehmigten Qualitätsmanagementsystems oder der Liste der unter dieses System fallenden KI-Systeme.

Die notifizierte Stelle prüft die vorgeschlagenen Änderungen und entscheidet, ob das geänderte Qualitätsmanagementsystem die in Nummer 3.2 genannten Anforderungen weiterhin erfüllt oder ob eine erneute Bewertung erforderlich ist.

Die notifizierte Stelle teilt dem Anbieter ihre Entscheidung mit. Die Mitteilung enthält die Ergebnisse der Prüfung der Änderungen und die begründete Bewertungsentscheidung.

4. Kontrolle der technischen Dokumentation

4.1. Zusätzlich zu dem in Nummer 3 genannten Antrag stellt der Anbieter bei der notifizierten Stelle seiner Wahl einen Antrag auf Bewertung der technischen Dokumentation für das KI-System, das er in Verkehr zu bringen oder in Betrieb zu nehmen beabsichtigt und das unter das in Nummer 3 genannte Qualitätsmanagementsystem fällt.

4.2. Der Antrag enthält:

a) den Namen und die Anschrift des Anbieters,

b) eine schriftliche Erklärung, dass derselbe Antrag bei keiner anderen notifizierten Stelle eingereicht worden ist,

c) die in Anhang IV genannte technische Dokumentation.

4.3. Die technische Dokumentation wird von der notifizierten Stelle geprüft. Dazu erhält die notifizierte Stelle, sofern dies relevant ist und beschränkt auf das zur Wahrnehmung der Aufgaben dieser Behörden erforderliche Maß, uneingeschränkten Zugang zu den verwendeten Trainings-, Validierungs- und Testdatensätzen, einschließlich, sofern dies relevant ist und im Rahmen der Sicherheitsmaßnahmen, über die Anwendungsprogrammierschnittstellen (API) oder andere einschlägige technische Mittel und Tools, die den Fernzugriff ermöglichen.

4.4. Bei der Prüfung der technischen Dokumentation kann die notifizierte Stelle vom Anbieter weitere Nachweise verlangen oder weitere Tests durch-

führen, um eine ordnungsgemäße Bewertung der Konformität des KI-Systems mit den Anforderungen in Titel III Kapitel 2 zu ermöglichen. Ist die notifizierte Stelle mit den vom Anbieter durchgeführten Tests nicht zufrieden, so führt sie gegebenenfalls unmittelbar selbst angemessene Tests durch.

4.5. Zum Quellcode des KI-Systems erhalten notifizierte Stellen auf begründete Anfrage und nur dann Zugang, wenn die folgenden kumulativen Bedingungen erfüllt sind:
   a) Der Zugang zum Quellcode ist zur Bewertung der Konformität des Hochrisiko-KI-Systems mit den in Titel III Kapitel 2 festgelegten Anforderungen notwendig, und
   b) die Test-/Prüfverfahren und Überprüfungen aufgrund der vom Anbieter bereitgestellten Daten und Dokumentation wurden ausgeschöpft oder haben sich als unzureichend erwiesen.

4.6. Die Entscheidung wird dem Anbieter oder dessen Bevollmächtigten mitgeteilt. Die Mitteilung enthält die Ergebnisse der Bewertung der technischen Dokumentation und die begründete Bewertungsentscheidung.
Erfüllt das KI-System die Anforderungen in Titel III Kapitel 2, so stellt die notifizierte Stelle eine EU-Bescheinigung über die Bewertung der technischen Dokumentation aus. Diese Bescheinigung enthält den Namen und die Anschrift des Anbieters, die Ergebnisse der Prüfung, etwaige Bedingungen für ihre Gültigkeit und die für die Identifizierung des KI-Systems notwendigen Daten.
Die Bescheinigung und ihre Anhänge enthalten alle zweckdienlichen Angaben für die Beurteilung der Konformität des KI-Systems und gegebenenfalls für die Kontrolle des KI-Systems während seiner Verwendung.
Entspricht das KI-System nicht den Anforderungen in Titel III Kapitel 2, so verweigert die notifizierte Stelle die Ausstellung einer EU-Bescheinigung über die Bewertung der technischen Dokumentation und unterrichtet den Antragsteller darüber, wobei sie ihre Weigerung ausführlich begründet.
Erfüllt das KI-System nicht die Anforderung in Bezug auf seine verwendeten Trainingsdaten, so muss das KI-System vor der Beantragung einer neuen Konformitätsbewertung erneut trainiert werden. In diesem Fall enthält die begründete Bewertungsentscheidung der notifizierten Stelle, mit der die Ausstellung der EU-Bescheinigung über die Bewertung der technischen Dokumentation verweigert wird, besondere Erläuterungen zu den

zum Trainieren des KI-Systems verwendeten Qualitätsdaten und insbesondere zu den Gründen für die Nichtkonformität.

4.7. Jede Änderung des KI-Systems, die sich auf die Konformität des KI-Systems mit den Anforderungen oder auf seine Zweckbestimmung auswirken könnte, bedarf der Genehmigung der notifizierten Stelle, die die EU-Bescheinigung über die Bewertung der technischen Dokumentation ausgestellt hat. Der Anbieter unterrichtet die notifizierte Stelle über seine Absicht, die oben genannten Änderungen vorzunehmen, oder wenn er auf andere Weise Kenntnis vom Eintreten solcher Änderungen erhält. Die notifizierte Stelle bewertet die beabsichtigten Änderungen und entscheidet, ob diese Änderungen eine neue Konformitätsbewertung gemäß Artikel 43 Absatz 4 erforderlich machen oder ob ein Nachtrag zu der EU-Bescheinigung über die Bewertung der technischen Dokumentation ausgestellt werden könnte. In letzterem Fall bewertet die notifizierte Stelle die beabsichtigten Änderungen, teilt dem Anbieter ihre Entscheidung mit und stellt ihm, sofern die Änderungen genehmigt wurden, einen Nachtrag zu der EU-Bescheinigung über die Bewertung der technischen Dokumentation aus.

5. Überwachung des genehmigten Qualitätsmanagementsystems

5.1. Mit der in Nummer 3 genannten Überwachung durch die notifizierte Stelle soll sichergestellt werden, dass der Anbieter die Anforderungen und Bedingungen des genehmigten Qualitätsmanagementsystems ordnungsgemäß einhält.

5.2. Zu Bewertungszwecken gewährt der Anbieter der notifizierten Stelle Zugang zu den Räumlichkeiten, in denen die Konzeption, die Entwicklung und das Testen der KI-Systeme stattfindet. Außerdem übermittelt der Anbieter der notifizierten Stelle alle erforderlichen Informationen.

5.3. Die notifizierte Stelle führt regelmäßig Audits durch, um sicherzustellen, dass der Anbieter das Qualitätsmanagementsystem pflegt und anwendet, und übermittelt ihm einen entsprechenden Prüfbericht. Im Rahmen dieser Audits kann die notifizierte Stelle die KI-Systeme, für die eine EU-Bescheinigung über die Bewertung der technischen Dokumentation ausgestellt wurde, zusätzlichen Tests unterziehen.

Literaturverzeichnis

Adadi, Amina/ Berrada, Mohammed, Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI), IEEE Access, 2018, S. 52138–52160.

Alexander, Christian, Gegenstand, Inhalt und Umfang des Schutzes von Geschäftsgeheimnissen nach der Richtlinie (EU) 2016/943, Wettbewerb in Recht und Praxis, 2017, S. 1034–1045.

- Geschäftsgeheimnisse und Ranking-Transparenz, Multimedia und Recht, 2021, S. 690–695.

Andrews, Robert/ Diederich, Joachim/ Tickle, Alan B., Survey and critique of techniques for extracting rules from trained artificial neural networks, Knowledge-Based Systems, 1995, S. 373–389.

Angwin, Julia/ Larson, Jeff/ Mattu, Surya/ Kirchner, Lauren, Machine Bias, ProPublica, https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing (zuletzt abgerufen am 26.10.2023).

Ann, Christoph, Geheimnisschutz – Kernaufgabe des Informationsmanagements im Unternehmen, Gewerblicher Rechtsschutz und Urheberrecht, 2014, S. 12–16.

Apel, Simon/ Kaulartz, Markus, Rechtlicher Schutz von Machine Learning-Modellen, Recht Digital, 2020, S. 24–34.

Artikel-29-Datenschutzgruppe, Leitlinien für Transparenz gemäß der Verordnung 2016/679, WP 260 rev.01.

Ateniese, Giuseppe/ Felici, Giovanni/ Mancini, Luigi V./ Spognardi, Angelo/ Villani, Antonio/ Vitali, Domenico, Hacking Smart Machines with Smarter Ones: How to Extract Meaningful Data from Machine Learning Classifiers, arXiv:1306.4447, 2013.

Baase, Sara, A gift of fire: social, legal, and ethical issues for computing technology, 4th ed, Upper Saddle River, NJ 2013.

Bach, Sebastian/ Binder, Alexander/ Montavon, Grégoire/ Klauschen, Frederick/ Müller, Klaus-Robert/ Samek, Wojciech, On Pixel-Wise Explanations for Non-Linear Classifier Decisions by Layer-Wise Relevance Propagation, PLoS ONE, 2015, S. 1–46.

Benkard, Georg (Hrsg.), Patentgesetz, Gebrauchsmustergesetz, Patentkostengesetz, 11., neubearbeitete Auflage, München 2015 (zitiert: *Bearb.*, in: Benkard, Patentgesetz).

Benkard, Georg/ Ehlers, Jochen/ Kinkeldey, Ursula (Hrsg.), Europäisches Patentübereinkommen, 3. Auflage, München 2019 (zitiert: *Bearb.*, in: Benkard, EPÜ).

Bermúdez, José Luis, Cognitive science: an introduction to the science of the mind, Third edition, Cambridge 2020.

Bischof, Elke/ Intveen, Michael, Einsatz künstlicher Intelligenz durch Unternehmen. Allgemeine Beschreibung und Fragen des Einsatzes insb. in der Automobilindustrie, IT-Rechtsberater 2019, S. 134–140.

Bundesregierung, Entwurf eines Gesetzes zur Umsetzung der Richtlinie (EU) 2016/943 zum Schutz von Geschäftsgeheimnissen vor rechtswidrigem Erwerb sowie rechtswidriger Nutzung und Offenlegung, BT-Drs. 19/4724.

- Strategie Künstliche Intelligenz der Bundesregierung, November 2018.

Burrell, Jenna, How the machine 'thinks': Understanding opacity in machine learning algorithms, Big Data & Society, 2016, S. 1–12.

Castelvecchi, Davide, The Black Box of AI, Nature, 2016, S. 20–23.

Craven, Mark W., Extracting comprehensible models from trained neural networks, PhD Thesis, Computer Science Department, University of Wisconsin, Madison, WI, 1996.

Deutsch, Florian/ Eggendorfer, Tobias, IT-Sicherheit, in: Taeger, Jürgen/ Pohle, Jan (Hrsg.), Computerrechts-Handbuch: Computertechnologie in der Rechts- und Wirtschaftspraxis, München 2021.

Deutscher Bundestag, Bericht der Enquete-Kommission Künstliche Intelligenz – Gesellschaftliche Verantwortung und wirtschaftliche, soziale und ökologische Potenziale, BT-Drs. 19/23700.

Dewes, Andreas, Lokale Nachvollziehbarkeit von ML-Modellen, in: Bitkom (Hrsg.), Blick in die Blackbox. Nachvollziehbarkeit von KI-Algorithmen in der Praxis, 2019, S. 22-31.

Doshi-Velez, Finale/Kim, Been, Towards A Rigorous Science of Interpretable Machine Learning, arXiv:1702.08608, 2017, S. 1–13.

Dreier, Thomas/Schulze, Gernot (Hrsg.), Urheberrechtsgesetz Kommentar, 7. Auflage, München 2022 (zitiert: *Bearb.*, in: Dreier/Schulze: UrhG).

Drexl, Josef/ Hilty, Reto/ Beneke, Francisco/ Desaunettes, Luc/ Finck, Michèle/ Globocnik, Jure/ Otero, Begoña Gonzalez/ Hoffmann, Jörg/ Hollander, Leonard/ Kim, Daria/ Richter, Heiko/ Scheuerer, Stefan/ Slowinski, Peter R./ Thonemann, Jannick, Technical Aspects of Artificial Intelligence: An Understanding from an Intellectual Property Law Perspective, Max Planck Institute for Innovation and Competition Research Paper No. 19-13, 2019.

Drexl, Josef/ Hilty, Reto/ Desaunettes-Barbero, Luc/ Globocnik, Jure/ Gonzalez Otero, Begoña/ Hoffmann, Jörg/ Kim, Daria/ Kulhari, Shraddha/ Richter, Heiko/ Scheuerer, Stefan/ Slowinski, Peter R./ Wiedemann, Klaus, Artificial Intelligence and Intellectual Property Law - Position Statement of the Max Planck Institute for Innovation and Competition of 9 April 2021 on the Current Debate, Max Planck Institute for Innovation and Competition Research Paper No. 21-10, 2021.

Dreyer, Stephan/Schulz, Wolfgang, Was bringt die Datenschutz-Grundverordnung für automatisierte Entscheidungssysteme?: Potenziale und Grenzen der Absicherung individueller, gruppenbezogener und gesellschaftlicher Interessen, 2018.

Ebeling, Werner/Freund, Jan/Schweitzer, Frank, Komplexe Strukturen: Entropie und Information, Wiesbaden 1998.

Ebers, Martin/ Hoch, Veronica R. S./ Rosenkranz, Martin/ Ruschemeier, Hannah/ Steinrötter, Björn, Der Entwurf für eine EU-KI-Verordnung: Richtige Richtung mit Optimierungsbedarf. Eine kritische Bewertung durch Mitglieder der Robotics & AI Law Society (RAILS), Recht Digital, 2021, S. 528–537.

Ebert, Andreas/ Spiecker gen. Döhmann, Indra, Der Kommissionsentwurf für eine KI-Verordnung der EU. Die EU als Trendsetter weltweiter KI-Regulierung, Neue Zeitschrift für Verwaltungsrecht, 2021, S. 1188–1193.

Edwards, Lilian/ Veale, Michael, Slave to the Algorithm? Why a „right to an explanation" is probably not the remedy you are looking for, Duke Law & Technology Review, 2017, S. 18–84.

Ebinger, Patrick/ Stiemerling, Oliver, Die urheberrechtliche Schutzfähigkeit von Künstlicher Intelligenz am Beispiel von Neuronalen Netzen, Computer und Recht, 2018, S. 761–770.

Ehsan, Upol/ Riedl, Mark O., Human-Centered Explainable AI: Towards a Reflective Sociotechnical Approach, in: Stephanidis, Constantine/ Kurosu, Masaaki/ Degen, Helmut/ Reinerman-Jones, Lauren (Hrsg.), HCI International 2020 - Late Breaking Papers: Multimodality and Intelligence, Cham 2020, S. 449–466.

Ertel, Wolfgang, Grundkurs Künstliche Intelligenz: eine praxisorientierte Einführung, 4., überarbeitete Auflage, Wiesbaden 2016.

Europäische Kommission, Vorschlag für eine Verordnung des Europäischen Parlaments und des Rates zur Festlegung harmonisierter Vorschriften für Künstliche Intelligenz (Gesetz über Künstliche Intelligenz) und zur Änderung bestimmter Rechtsakte der Union, COM(2021) 206 final (zitiert: *Europäische Kommission*, Entwurf der KI-Verordnung, COM(2021) 206 final).

Europäische Kommission, Weißbuch zur Künstlichen Intelligenz – ein europäisches Konzept für Exzellenz und Vertrauen, COM(2020) 65 final.

Europäisches Patentamt, Richtlinien für die Prüfung im Europäischen Patentamt, München 2021.

Flasiński, Mariusz, Introduction to Artificial Intelligence, Cham 2016.

Fraunhofer-Institut für Nachrichtentechnik, Explainable AI Demos, https://lrpserver.hhi.fraunhofer.de/image-classification (zuletzt abgerufen am 26.10.2023).

Fredrikson, Matt/ Jha, Somesh/ Ristenpart, Thomas, Model Inversion Attacks that Exploit Confidence Information and Basic Countermeasures, in: Ray, Indrajit/ Li, Ninghui/ Kruegel, Christopher (Hrsg.), Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security, 2015.

Geitz, Eckhard/ Vater, Christian/ Zimmer-Merkle, Silke, Black Boxes - Versiegelungskontexte und Öffnungsversuche: interdisziplinäre Perspektiven, Berlin, Boston 2020.

Gesellschaft für Informatik, Technische und rechtliche Betrachtungen algorithmischer Entscheidungsverfahren. Gutachten der Fachgruppe Rechtsinformatik der Gesellschaft für Informatik e.V. im Auftrag des Sachverständigenrats für Verbraucherfragen, Berlin 2018.

Ghahramani, Zoubin, Probabilistic machine learning and artificial intelligence, Nature, 2015, S. 452–459.

Gierschmann, Sibylle/ Schlender, Katharina/ Stentzel, Rainer/ Veil, Winfried, Kommentar Datenschutz-Grundverordnung, Bonn 2018 (zitiert: *Bearb.*, in: Gierschmann DSGVO).

Gilpin, Leilani H./ Bau, David/ Yuan, Ben Z./ Bajwa, Ayesha/ Specter, Michael/ Kagal, Lalana, Explaining Explanations: An Overview of Interpretability of Machine Learning, arXiv:180600069, 2019, S. 1–7.

Gilpin, Leilani H./ Testart, Cecilia/ Fruchter, Nathaniel/ Adebayo, Julius, Explaining Explanations to Society, arXiv:1901.06560v1, 2019, S. 1–7.

Gola, Peter (Hrsg.), Datenschutz-Grundverordnung: Kommentar, 2. Auflage, München 2018 (zitiert: *Bearb.*, in: Gola DS-GVO).

Golland, Alexander, Dynamic Pricing: Algorithmen zwischen Ökonomie und Datenschutz, in: Taeger, Jürgen (Hrsg.), Die Macht der Daten und der Algorithmen: Regulierung von IT, IoT und KI, Edewecht 2019, S. 61–76.

Grützmacher, Malte, Urheber-, Leistungs- und Sui-generis-Schutz von Datenbanken. Eine Untersuchung des europäischen, deutschen und britischen Rechts, 1999.

Guidotti, Riccardo/ Monreale, Anna/ Giannotti, Fosca/ Pedreschi, Dino/ Ruggieri, Salvatore/ Turini, Franco, Factual and Counterfactual Explanations for Black Box Decision Making, IEEE Intelligent Systems, 2019, S. 14–23.

Guidotti, Riccardo/ Monreale, Anna/ Ruggieri, Salvatore/ Turini, Franco/Giannotti, Fosca/ Pedreschi, Dino, A Survey of Methods for Explaining Black Box Models, ACM Computing Surveys, 2019, S. 1–42.

Hacker, Philipp, Europäische und nationale Regulierung von Künstlicher Intelligenz, Neue Juristische Wochenschrift, 2020, S. 2142–2147.

Hacker, Philipp/ Krestel, Ralf/ Grundmann, Stefan/ Naumann, Felix, Explainable AI under contract and tort law: legal incentives and technical challenges, Artificial Intelligence and Law, 2020, S. 415–439.

Harley, Adam W., An Interactive Node-Link Visualization of Convolutional Neural Networks, https://adamharley.com/nn_vis/mlp/3d.html (zuletzt abgerufen am 26.10.2023).

Harte-Bavendamm, Henning/ Ohly, Ansgar/ Kalbfuß, Björn (Hrsg.), Gesetz zum Schutz von Geschäftsgeheimnissen: Kommentar, 1. Auflage, München 2020 (zitiert: *Bearb.*, in: Harte-Bavendamm/Ohly/Kalbfus, GeschGehG).

Hartmann, Frank/Prinz, Matthias, Immaterialgüterrechtlicher Schutz von Systemen Künstlicher Intelligenz, in: Taeger, Jürgen (Hrsg.), Rechtsfragen digitaler Transformationen - Gestaltung digitaler Veränderungsprozesse durch Recht, Edewecht 2018, S. 769–789.

Hauck, Ronny, Geheimnisschutz im Zivilprozess – was bringt die neue EU-Richtlinie für das deutsche Recht?, Neue Juristische Wochenschrift, S. 2218–2223.

Hauck, Ronny/Cevc, Baltasar, Patentschutz für Systeme Künstlicher Intelligenz?, Zeitschrift für geistiges Eigentum, 2019, S. 135–169.

Haykin, Simon, Neural Networks: A Comprehensive Foundation, 2. Edition, Upper Saddle River [u.a.] 2009.

Heermann, Peter W./ Schlingloff, Jochen (Hrsg.), Münchener Kommentar zum Lauterkeitsrecht. Band 2: Besondere Fallgruppen und Rechtsgebiete: §§ 7a-20 UWG, 3. Auflage, München 2022 (zitiert: *Bearb.*, in: MüKo Lauterkeitsrecht, Band 2).

Hendricks, Lisa Anne/ Akata, Zeynep/ Rohrbach, Marcus/ Donahue, Jeff/ Schiele, Bernt/ Darrell, Trevor, Generating Visual Explanations, in: Leibe, Bastian/ Matas, Jiri/ Sebe, Nicu/ Welling, Max (Hrsg.), Computer Vision – ECCV 2016, Cham 2016, S. 3–19.

Hessel, Stefan/Lesser, Lena, Rechtlicher Schutz maschinengenerierter Daten, Multimedia und Recht, 2020, S. 647–650.

Hochrangige Expertengruppe für Künstliche Intelligenz, eingesetzt von der Europäischen Kommission im Juni 2018: Eine Definition der KI: wichtigste Fähigkeiten und Wissenschaftsgebiete.

Hochrangige Expertengruppe für Künstliche Intelligenz, eingesetzt von der Europäischen Kommission im Juni 2018, Ethik-Leitlinien für eine vertrauenswürdige KI.

Hoeren, Thomas/ Wehkamp, Nils, Individualität im Quellcode? Softwareschutz und Urheberrecht, Computer und Recht, 2018, S. 1–7.

Hohman, Fred/ Kahng, Minsuk/ Pienta, Robert/ Chau, Duen Horng, Visual Analytics in Deep Learning: An Interrogative Survey for the Next Frontiers, IEEE Transactions on Visualization and Computer Graphics, 2019, S. 2674–2693.

Holzinger, Andreas, Explainable AI (ex-AI), Informatik-Spektrum, 2018, S. 138–143.

Hoppe, Daniel, in: Hoppe, Daniel/ Oldekop, Axel (Hrsg.), Geschäftsgeheimnisse. Schutz von Know-how und Geschäftsinformationen. Praktikerhandbuch mit Mustern, 2020.

Hornung, Oliver, Die EU-Datenbank-Richtlinie und ihre Umsetzung in das deutsche Recht: eine Untersuchung unter besonderer Berücksichtigung des Schutzrechts sui generis nach der EU-Datenbank-Richtlinie, 1. Auflage, Baden-Baden 1998.

Horowitz, Ellis/ Sahni, Sartaj, Algorithmen: Entwurf und Analyse, Berlin 1981.

Huber, Bertram, Faktischer Know-how-Schutz in der Unternehmenspraxis, in: Ann, Christoph/ Loschelder, Michael/ Grosch, Marcus (Hrsg.), Praxishandbuch Know-how-Schutz, Köln 2010.

Kalbfus, Björn, Angemessene Geheimhaltungsmaßnahmen nach der Geschäftsgeheimnis-Richtlinie, Gewerblicher Rechtsschutz und Urheberrecht in der Praxis, 2017, S. 391–393.

- Die neuere Rechtsprechung des BGH zum Schutz von Betriebs- und Geschäftsgeheimnissen, Wettbewerb in Recht und Praxis, 2013, S. 584–590.

Kaminski, Margot E., The Right to Explanation, Explained, Berkely Tech Law Journal, 2019, S. 190–218.

Kaplan, Jerry, Künstliche Intelligenz, 1. Auflage, Frechen 2017.

Knight, Will, The dark secret at the heart of AI, Technology review (1998), 2017, S. 53-63.

Köhler, Helmut/ Bornkamm, Joachim/ Feddersen, Björn/ Alexander, Christian (Hrsg.), Gesetz gegen den unlauteren Wettbewerb: GeschGehG, PAngV, UKlaG, DL-InfoV, 39., neu bearbeitete Auflage, München 2021 (zitiert: *Bearb.*, in: Köhler/Bornkamm/Feddersen, GeschGehG).

Koch, Raphael/ Biggen, Christine, Der Einsatz Künstlicher Intelligenz zur Organisation und proaktiven Überprüfung von Onlinebewertungen, Neue Juristische Wochenschrift, 2020, S. 2921–2925.

Konertz, Roman/ Schönhof, Raoul, Das technische Phänomen „Künstliche Intelligenz" im allgemeinen Zivilrecht: Eine kritische Betrachtung im Lichte von Autonomie, Determinismus und Vorhersehbarkeit, 2020.

Kruse, Rudolf/ Borgelt, Christian/Braune, Christian/ Klawonn, Frank/ Moewes, Christian/ Steinbrecher, Matthias, Computational Intelligence: Eine methodische Einführung in Künstliche Neuronale Netze, Evolutionäre Algorithmen, Fuzzy-Systeme und Bayes-Netze, Wiesbaden 2015.

Kühling, Jürgen/ Buchner, Benedikt/ Bäcker, Matthias (Hrsg.), Datenschutz-Grundverordnung/BDSG: Kommentar, 3. Auflage, München 2020 (zitiert: *Bearb.*, in: Kühling/Buchner DS-GVO).

Kumkar, Lea Katharina/ Roth-Isigkeit, David, Erklärungspflichten bei automatisierten Datenverarbeitungen nach der DSGVO, JuristenZeitung, 2020, S. 277–286.

Kuß, Christian/ Sassenberg, Thomas, in: Sassenberg, Thomas/ Faber, Tobias (Hrsg.), Rechtshandbuch Industrie 4.0 und Internet of Things: Praxisfragen und Perspektiven der digitalen Zukunft, 2. Auflage, München 2020.

Lämmel, Uwe/ Cleve, Jürgen, Künstliche Intelligenz, Wissensverarbeitung - Neuronale Netze, 2020.

Lapuschkin, Sebastian/ Wäldchen, Stephan/ Binder, Alexander/ Montavon, Grégoire/ Samek, Wojciech/ Müller, Klaus-Robert, Unmasking Clever Hans predictors and assessing what machines really learn, Nature Communications, 2019, S. 1–8.

LeCun, Yann/ Bengio, Yoshua/ Hinton, Geoffrey, Deep learning, Nature, 2015, S. 436–444.

Lipton, Zachary C., The Mythos of Model Interpretability, ICML Workshop on Human Interpretability in Machine Learning (WHI 2016), 2016, S. 96–100.

Loewenheim, Ulrich/ Leistner, Matthias/ Ohly, Ansgar (Hrsg.), Urheberrecht: UrhG, KUG, VGG: Kommentar, 6. Auflage, München 2020 (zitiert: *Bearb.*, in: Schricker/Loewenheim, UrhG).

Lundberg, Scott, SHAP Documentation, https://shap.readthedocs.io/en/latest/api_examples.html#plots (zuletzt abgerufen am 26.10.2023).

Lundberg, Scott/ Lee, Su-In, A Unified Approach to Interpreting Model Predictions, 31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA, 2017, S. 1–10.

Lyre, Holger, Informationstheorie. Eine philosophisch-naturwissenschaftliche Einführung, München 2002.

Maaßen, Stefan, „Angemessene Geheimhaltungsmaßnahmen" für Geschäftsgeheimnisse, Gewerblicher Rechtsschutz und Urheberrecht, 2019, S. 352–260.

Mainzer, Klaus, Künstliche Intelligenz – Wann übernehmen die Maschinen?, Berlin, Heidelberg 2019.

Malgieri, Gianclaudio/ Comandé, Giovanni, Why a Right to Legibility of Automated Decision-Making Exists in the General Data Protection Regulation, International Data Privacy Law, 2017, S. 243–265.

Martini, Mario, Blackbox Algorithmus – Grundfragen einer Regulierung Künstlicher Intelligenz, Berlin, Heidelberg 2019.

Ménière, Yann/ Pihlajamaa, Heli, Künstliche Intelligenz in der Praxis des EPA, Gewerblicher Rechtsschutz und Urheberrecht, 2019, S. 332–336.

Misheva, Branka Hadji/ Osterrieder, Joerg/ Hirsa, Ali/ Kulkarni, Onkar/ Lin, Stephen Fung, Explainable AI in Credit Risk Management, arXiv:2103.00949, 2021, S. 1–16.

Mitchell, Tom M., Machine Learning, New York 1997.

Müllmann, Dirk, Auswirkungen der Industrie 4.0 auf den Schutz von Betriebs- und Geschäftsgeheimnissen, Wettbewerb in Recht und Praxis, 2018, S. 1177–1182.

Nägele, Thomas/ Apel, Simon, KI und Urheberrecht, in: Kaulartz, Markus/ Braegelmann, Tom (Hrsg.), Rechtshandbuch Artificial Intelligence und Machine Learning, München, 2020.

Nägerl, Joel/ Neuburger, Benedikt/ Steinbach, Frank, Künstliche Intelligenz: Paradigmenwechsel im Patentsystem, Gewerblicher Rechtsschutz und Urheberrecht, 2019, S. 336–341.

Nebel, Jens/ Stiemerling, Oliver, Aktuelle Programmiertechniken und ihr Schutz durch § 69a UrhG, Computer und Recht, 2016, S. 61–69.

Newell, Allen/ Simon, Herbert A., Computer Science as Empirical Inquiry: Symbols and Search, Communications of the ACM, 1975, S. 113–126.

Nguyen, Anh/ Yosinski, Jason/ Clune, Jeff, Understanding Neural Networks via Feature Visualization: A Survey, in: Samek, Wojciech/ Montavon, Grégoire/ Vedaldi, Andrea/ Hansen, Lars Kai/ Müller, Klaus-Robert (Hrsg.), Explainable AI: Interpreting, Explaining and Visualizing Deep Learning, Cham 2019, S. 55-76.

Oh, Seong Joon/ Schiele, Bernt/ Fritz, Mario, Towards Reverse-Engineering Black-Box Neural Networks, in: Samek, Wojciech/ Montavon, Grégoire/ Vedaldi, Andrea/ Hansen, Lars Kai/ Müller, Klaus-Robert (Hrsg.), Explainable AI: Interpreting, Explaining and Visualizing Deep Learning, Cham 2019, 121-144.

Ohly, Ansgar, Das neue Geschäftsgeheimnisgesetz im Überblick, Gewerblicher Rechtsschutz und Urheberrecht, 2019, S. 441–451.

Olden, Julian D./ Jackson, Donald A., Illuminating the „black box“: a randomization approach for understanding variable contributions in artificial neural networks, Ecological Modelling, 2002, S. 135–150.

Paal, Boris/ Pauly, Daniel (Hrsg.), Datenschutz-Grundverordnung Bundesdatenschutzgesetz, München 2021 (zitiert: *Bearb.*, in: Paal/Pauly DS-GVO).

Papernot, Nicolas/ McDaniel, Patrick/ Goodfellow, Ian/ Jha, Somesh/ Celik, Z. Berkay/ Swami, Ananthram, Practical Black-Box Attacks against Machine Learning, Proceedings of the 2017 ACM Asia Conference on Computer and Communications Security, Abu Dhabi, UAE, 2017, S. 506–519.

Pasquale, Frank, The black box society: the secret algorithms that control money and information, Cambridge 2015.

Rat der Europäischen Union, Vorschlag für eine Verordnung des Europäischen Parlaments und des Rates zur Festlegung harmonisierter Vorschriften für Künstliche Intelligenz (Gesetz über Künstliche Intelligenz) und zur Änderung bestimmter Rechtsakte der Union, Allgemeine Ausrichtung vom 6.12.2022, 15698/22 (zitiert: Entwurf der KI-Verordnung, Allgemeine Ausrichtung).

Redeker, Helmut, IT-Recht, 7., neubearbeitete Auflage, München 2020.

Ribeiro, Marco Tulio/ Singh, Sameer/ Guestrin, Carlos, „Why Should I Trust You?“: Explaining the Predictions of Any Classifier, Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016, S. 1135–1144.

Rich, Elaine, Artificial intelligence, New York 1983.

Roos, Philipp/ Weitz, Caspar Alexander, Hochrisiko-KI-Systeme im Kommissionsentwurf für eine KI-Verordnung. IT- und produktsicherheitsrechtliche Pflichten von Anbietern, Einführern, Händlern und Nutzern, Multimedia und Recht, 2021, S. 844–851.

Roßnagel, Alexander/ Nebel, Maxi/ Richter, Philipp, Was bleibt vom Europäischen Datenschutzrecht? Überlegungen zum Ratsentwurf der DS-GVO, Zeitschrift für Datenschutz, 2015, S. 455–460.

Rudin, Cynthia, Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead, Nature Machine Intelligence, 2019, S. 206–215.

Russell, Stuart J./ Norvig, Peter, Artificial intelligence: a modern approach, 4. Auflage, Hoboken 2021.

Sagstetter, Thomas, Big Data und der europäische Rechtsrahmen: Status quo und Reformbedarf im Lichte der Trade-Secrets-Richtlinie 2016/943/EU, in: Maute, Lena/Mackenrodt, Mark-Oliver (Hrsg.), Recht als Infrastruktur für Innovation, München 2018.

Samek, Wojciech/ Müller, Klaus-Robert, Towards Explainable Artificial Intelligence, in: Samek, Wojciech/ Montavon, Grégoire/ Vedaldi, Andrea/ Hansen, Lars Kai/ Müller, Klaus-Robert (Hrsg.), Explainable AI: Interpreting, Explaining and Visualizing Deep Learning, Cham 2019, S. 5-22.

Schaaf, Nina/ Huber, Marco, Extraktion von Erklärungen zu Produktionsprozessen aus künstlichen Neuronalen Netzen, in: Bitkom (Hrsg.), Blick in die Blackbox. Nachvollziehbarkeit von KI-Algorithmen in der Praxis, Berlin 2019.

Scheja, Katharina, Schutz von Algorithmen in Big Data Anwendungen, Computer und Recht, 2018, S. 485–492.

Schmidt, Marcus, Handbuch IT- und Datenschutzrecht, 3. Auflage, München 2019.

Schönfeld, Dagmar/ Klimant, Herbert/ Piotraschke, Rudi, Informations- und Kodierungstheorie, Wiesbaden 2012.

Selbst, Andrew D./ Powles, Julia, Meaningful information and the right to explanation, International Data Privacy Law, 2017, S. 233–242.

Shokri, Reza/ Stronati, Marco/ Song, Congzheng/ Shmatikov, Vitaly, Membership Inference Attacks against Machine Learning Models, 2017, S. 1–16.

Simitis, Spiros/ Hornung, Gerrit/ Spiecker Döhmann, Indra (Hrsg.), Datenschutzrecht: DSGVO mit BDSG, 1. Auflage, Baden-Baden 2019 (zitiert: *Bearb.*, in: Simitis/Hornung/Spiecker Datenschutzrecht).

Söbbing, Thomas, Algorithmen und urheberrechtlicher Schutz, Computer und Recht, 2020, S. 223–228.

- Künstliche neuronale Netze. Rechtliche Betrachtung von Software- und KI-Lernstrukturen, Multimedia und Recht, 2021, S. 111–116.

Spindler, Gerald/ Schuster, Fabian/ Anton, Katharina (Hrsg.), Recht der elektronischen Medien: Kommentar, 4. Auflage, München 2019 (zitiert: *Bearb.*, in: Spindler/Schuster/Anton, Recht der elektronischen Medien).

Springorum, Harald, The protection of neural networks according to German and European law, in: Brunnstein, Klaus/ Sint, Peter Paul (Hrsg.), Intellectual Property Rights and New Technologies. Proceedings of the KnowRight'95 Conference, München 1995, S. 204-213

Surblyté, Gintaré, Data-Driven Economy and Artificial Intelligence: Emerging Competition Law Issues?, Wirtschaft und Wettbewerb, S. 120–127.

- Enhancing TRIPS: Trade Secrets and Reverse Engineering, in: Ullrich, Hanns/ Hilty, Reto/ Lamping, Matthias/ Drexl, Josef (Hrsg.), TRIPS plus 20: From Trade Rules to Market Principles, Berlin, Heidelberg 2016, S. 725-760

Sydow, Gernot (Hrsg.), Europäische Datenschutzgrundverordnung: Handkommentar, 2. Auflage, Baden-Baden 2018 (zitiert: *Bearb.*, in: Sydow DSGVO).

Taeger, Jürgen/ Gabel, Detlev/ Arning, Marian (Hrsg.), DSGVO - BDSG: Kommentar, 3., völlig neu bearbeitete und wesentlich erweiterte Auflage, Frankfurt am Main 2019 (zitiert: *Bearb.*, in: Taeger/Gabel DS-GVO).

Thomas, L. C./ Edelman, David B./ Crook, Jonathan N., Credit scoring and its applications, Second edition, Philadelphia 2017.

Tittmann, Peter, Graphentheorie: eine anwendungsorientierte Einführung, 3., aktualisierte Auflage, München 2019.

Tochtermann, Lea, Immaterialgüterrechtlicher Schutz von KI de lege ferenda, in: Kaulartz, Markus/Braegelmann, Tom (Hrsg.), Rechtshandbuch Artificial Intelligence und Machine Learning, München 2020, S. 322-335.

Tramèr, Florian/ Zhang, Fan/ Juels, Ari/ Reiter, Michael K./ Ristenpart, Thomas, Stealing Machine Learning Models via Prediction APIs, in: Holz, Thorsten/ Savage, Stefan (Hrsg.), Proceedings of the 25th USENIX Security Symposium August 10–12, 2016, S. 601-618.

Triebe, Christian, Reverse Engineering im Lichte des Urheber- und Geschäftsgeheimnisschutzes, Wettbewerb in Recht und Praxis, 2018, S. 795–805.

Vieth, Kilian/ Wagner, Ben, Teilhabe, ausgerechnet: Wie algorithmische Prozesse Teilhabechancen beeinflussen können, 2017.

Vigen, Tyler, Spurious correlations, https://www.tylervigen.com/spurious-correlations (zuletzt abgerufen am 26.10.2023).

Voosen, Paul, How AI detectives are cracking open the black box of deep learning, http://www.sciencemag.org/news/2017/07/how-ai-detectives-are-cracking-open-black-box-deep-learning (zuletzt abgerufen am 26.10.2023).

Wachter, Sandra/ Mittelstadt, Brent/ Floridi, Luciano, Why a Right to Explanation of Automated Decision-Making Does Not Exist in the General Data Protection Regulation, International Data Privacy Law, 2017, S. 76–99.

Wachter, Sandra/ Mittelstadt, Brent/Russell, Chris, Counterfactual Explanations Without Opening the Black Box: Automated Decisions and the GDPR, Harvard Journal of Law & Technology, 2018, S. 841–887.

Wandtke, Artur-Axel/ Bullinger, Winfried (Hrsg.), Praxiskommentar zum Urheberrecht, 5. Auflage, München 2019 (zitiert: *Bearb.*, in: Wandtke/Bullinger, UrhG).

Weller, Adrian, Transparency: Motivations and Challenges, in: Samek, Wojciech/ Montavon, Grégoire/ Vedaldi, Andrea/ Hansen, Lars Kai/ Müller, Klaus-Robert (Hrsg.), Explainable AI: Interpreting, Explaining and Visualizing Deep Learning, Cham 2019, S. 23-40.

West, David, Neural network credit scoring models, Computers & Operations Research, 2000, S. 1131–1152.

Wieder, Clemens, Datenschutzrechtliche Betroffenenrechte bei der Verarbeitung von personenbezogenen Daten mittels Künstlicher Intelligenz, in: Taeger, Jürgen (Hrsg.), Rechtsfra-

gen digitaler Transformationen - Gestaltung digitaler Veränderungsprozesse durch Recht, Edewecht 2018, S. 505–518.

Wiese, Elena, Die EU-Richtlinie über den Schutz vertraulichen Know-hows und vertraulicher Geschäftsinformationen, Berlin 2018.

Wischmeyer, Thomas, Regulierung intelligenter Systeme, Archiv des öffentlichen Rechts, 2018, S. 1–66.

Wissenschaftliche Dienste des Deutschen Bundestags, Künstliche Intelligenz und Machine Learning. Eine urheberrechtliche Betrachtung, WD 10 - 3000 - 67/18.

Yeh, Chih-Kuan/ Kim, Been/ Arik, Sercan O./ Li, Chun-Liang/ Pfister, Tomas/ Ravikumar, Pradeep, On Completeness-aware Concept-Based Explanations in Deep Neural Networks, arXiv:1910.07969v5, 2020, S. 1–27.

Yosinski, Jason/ Clune, Jeff/ Nguyen, Anh/ Fuchs, Thomas/ Lipson, Hod, Understanding Neural Networks Through Deep Visualization, Deep Leaning Workshop, 31st International Conference on Machine Learning, Lille, France, 2015, S. 1–12.

Zech, Herbert, „Industrie 4.0“ - Rechtsrahmen für eine Datenwirtschaft im digitalen Binnenmarkt, Gewerblicher Rechtsschutz und Urheberrecht, 2015, S. 1151–1160.

- A legal framework for a data economy in the European Digital Single Market: rights to use data, Journal of Intellectual Property Law & Practice, 2016, S. 460–470.
- Information als Schutzgegenstand, Tübingen 2012.
- Risiken Digitaler Systeme: Robotik, Lernfähigkeit und Vernetzung als aktuelle Herausforderungen für das Recht, Weizenbaum Series, 2020.

digital | recht

Schriften zum Immaterialgüter-, IT-,
Medien-, Daten- und Wettbewerbsrecht

Andrea Linhart

Information aus der Blackbox

Andrea Linhart untersucht, wie sich zunehmende Transparenzpflichten auf den Schutz von Künstlicher Intelligenz als Geschäftsgeheimnis auswirken. Speziell für trainierte Künstliche Neuronale Netze entwirft sie ein *System abgestufter Transparenz*, welches das Spannungsverhältnis zwischen Transparenz und Geheimnisschutz auflösen kann.