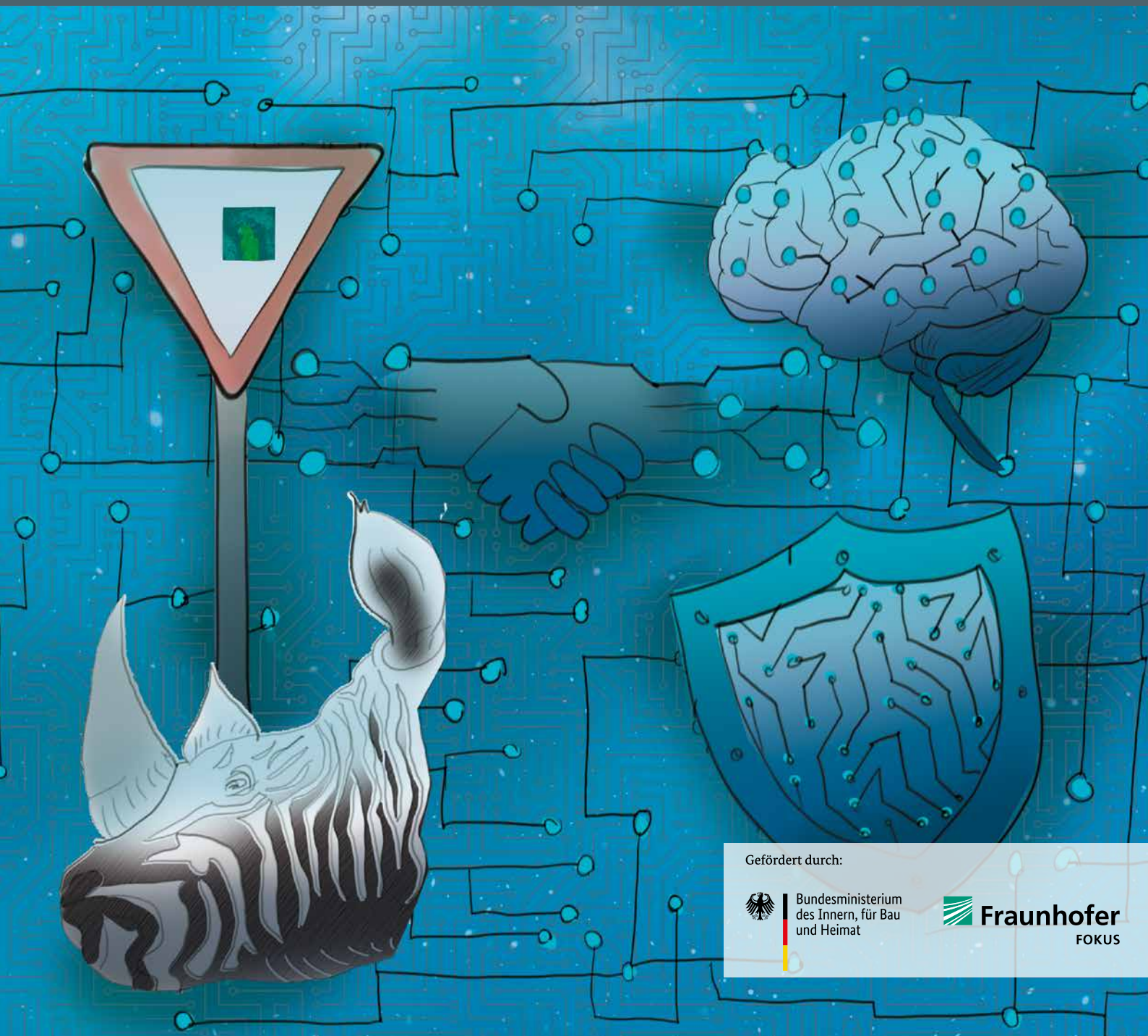
Kompetenzzentrum
Öffentliche IT

FORSCHUNG FÜR DEN DIGITALEN STAAT

INNOVATIONSFELDER ÖFFENTLICHER IT 2019/2020



Gefördert durch:

Bundesministerium des Innern, für Bau und Heimat

Fraunhofer FOKUS

IMPRESSUM

Autor:innen
Silke Cuno, Jan Gottschick, Dorian Grosch, Jan Dennis Gumz, Karoline Krenn, Nicole Opiela, Jens Tiemann, Mike Weber, Christian Welzel, Armin Wolf

Danksagung:
Wir danken Thilo Ernst, Franziska Frankenfeld, Maximilian Gahntz, Gabriele Goldacker, Simon Sebastian Hunt und Maximilian Kupi, die mit ihrem Fachwissen zu dieser Publikation beigetragen haben.

Gestaltung:
Reiko Kammer

Herausgeber:
Kompetenzzentrum Öffentliche IT
Fraunhofer-Institut für Offene Kommunikationssysteme FOKUS
Kaiserin-Augusta-Allee 31, 10589 Berlin
Telefon: +49-30-3463-7173
Telefax: +49-30-3463-99-7173
info@oeffentliche-it.de
www.oeffentliche-it.de
www.fokus.fraunhofer.de

ISBN: 978-3-9819921-7-5

1. Auflage Dezember 2019

Links:

Alle in dieser enthaltenen Weblinks wurden zuletzt am 30.11.2019 oder später abgerufen.

Bildnachweise:

Seite	Autor:innen	Quelle	Lizenz
1, 10, 16, 24, 32, 35, 40	Martha Friedrich		CC BY 3.0
21 (links)	Perlinator	pixabay.com	CC 0
21 (mitte)	MartinThoma	wikimedia.org	CC0 1.0
21 (rechts)	MartinThoma	wikimedia.org	CC0 1.0

VORWORT

»Sich sehnsüchtig der Zukunft erinnern« – das wollten wir mit der ersten Analyse[1] der Innovationsfelder öffentlicher IT. Nach sechs Jahren kontinuierlichen Trendmonitorings mit diversen Publikationen, beispielsweise Trendblätter[2], Technologiesonare[3] und Zukunftsszenarien[4], bleibt die Sehnsucht genauso groß.

Zukunftsbezogene Aussagen haben gegenüber solchen über die Vergangenheit die besondere Herausforderung, erst in der Zeit überprüfbar zu sein. Aber es geht bei Weitem nicht nur darum, die Zukunft treffsicher vorherzusagen. Vielmehr rekurrieren Zukunftsaussagen immer auf das hier und jetzt, in dem sie getroffen werden. So können sie die Entscheidungsgrundlage erweitern und den Akteur:innen mehr Zeit für notwendige Handlungen verschaffen. Zeichnen sich dystopische Entwicklungen und große neue Herausforderungen wie etwa bei den neuen Angriffsvektoren für KI-Anwendungen ab, dann muss es das Ziel jeder Zukunftsforschung sein, die eigene Vorausschau nicht eintreten zu lassen: Die Selbstzerstörung der eigenen Vorhersagen ist der größte Erfolg.

Diese zweite Analyse der Innovationsfelder öffentlicher IT zielt auf die Identifikation großer Entwicklungslinien. Im hochdynamischen Feld der IT braucht es dafür einen intensiven Blick auf die aktuellsten wissenschaftlichen Publikationen. Was heute an der Spitze der IT-Forschung diskutiert wird, kann binnen kurzer Zeit bereits den Markt beherrschen. Die Darstellung der Innovationsfelder richtet sich entsprechend nicht nur an technologisch Interessierte, die neue Anwendungsmöglichkeiten kennen lernen möchten. Die skizzierten Entwicklungen fordern mitunter auch Staat und Verwaltung heraus. Entsprechend werden Regulierungsaspekte genauso angesprochen, wie Herausforderungen der Anwendung in der eigenen Organisation.

Die bibliometrische Analyse von etwa zwei Millionen Konferenzbeiträgen hat fünf Innovationsfelder hervorgebracht. Künstliche Intelligenz in unterschiedlichen Fassetten dominiert die aktuelle wissenschaftliche Auseinandersetzung, ein Ergebnis, das auch durch das ÖFIT-Trendgebirge[5] gestützt wird. Die in dieser Publikation aufgezeigten Entwicklungslinien können dazu beitragen, die notwendige Diskussion über Möglichkeiten und Grenzen dieser und anderer Technologien auf eine informierte Basis zu stellen.

Dass dabei nicht immer jede technische Möglichkeit auch in die breite Anwendung kommt, zeigt die Rückschau auf die Ergebnisse der ersten Analyse der Innovationsfelder öffentlicher IT. Begriffe können sich wandeln, Konzepte werden weiterentwickelt, Tendenzen verstärken oder verlangsamen sich und mancher damals identifizierte Trend lässt sich in druckfrischen Trendreports wiederfinden. Dabei gilt damals wie heute, dass die aufgezeigten Möglichkeitsräume die Auseinandersetzung mit den bereits begonnenen Entwicklungen erleichtert.

Wir wünschen allen sich sehnsüchtig der Zukunft erinnernden Entscheider:innen und technisch Interessierten in Politik und Verwaltung eine anregende Lektüre.

Ihr Kompetenzzentrum Öffentliche IT

---

[1] Fromm, J.; Gauch, S.; Kaiser, T.; Weber, M. (2013).

[2] https://www.oeffentliche-it.de/trendschau.

[3] https://www.oeffentliche-it.de/trendsonar.

[4] Opiela, N.; Mohabbat Kar, R.; Thapa, B.; Weber, M. (2018).

[5] https://www.oeffentliche-it.de/trendgebirge.

»WE CAN ONLY SEE A SHORT DISTANCE AHEAD, BUT WE CAN SEE PLENTY THERE THAT NEEDS TO BE DONE.« ALAN TURING

INHALTSVERZEICHNIS

EINLEITUNG

Für diese Publikation wurden fünf Innovationsfelder[6] ermittelt, deren Beschreibung und Bewertung das Kernstück der vorliegenden Publikation darstellen:

1. Ressourceneffiziente KI:
   Fortschritte bei Software und Hardware, die performante KI abseits von Rechenzentren ermöglichen.

2. Künstlicher Realismus:
   Die durch innovative Algorithmen deutlich erleichterte Manipulation von Texten sowie Audio-, Bild- und Videomaterial mit täuschend echten Ergebnissen. Beispielhaft seien hier Deepfakes genannt.

3. Blockchain – Zwischen Hype und Wirklichkeit:
   Innovationen wie Smart Contracts und neue Konsensverfahren, die der Blockchain-Technologie den Weg zu weiteren praktischen Anwendungsmöglichkeiten ebnen.

4. Die Achillesferse der KI?:
   Durch neu entdeckte Schwachstellen und Angriffsmöglichkeiten wie beispielsweise Adversarial Examples, werden Fragen bezüglich der sinnvollen Einsatzgebiete von KI aufgeworfen.

5. Maschinen verstehen Menschen:
   Algorithmen, die menschliche Haltungen und Absichten einordnen können. Beispielsweise wird Hate Speech Detection eingesetzt, um die Verbreitung von Hassrede in sozialen Netzwerken einzudämmen.

In Kapitel 2 sind die wesentlichen Trends der fünf Innovationsfelder erläutert, mögliche Anwendungen und Auswirkungen dargestellt sowie Wege zur Gestaltung der zukünftigen Entwicklung aufgezeigt. Generell wurde ein starker Fokus auf die Datenauswertung und -visualisierung gelegt. Wie die in dieser Publikation verwendeten Grafiken zu lesen sind, ist in Kapitel 1 dargelegt.

Um die Innovationsfelder 2019/2020 zu identifizieren, wurden zunächst wissenschaftliche Publikation automatisiert nach vorher bestimmten Kriterien analysiert. Aus den sich ergebenden Themen wurden in Expert:innenworkshops Trends ermittelt, die dann zu Innovationsfeldern aggregiert wurden. Der Prozess ist ausführlich in Kapitel 4 beschrieben.

In Kapitel 3 sind die Innovationsfelder »Anything as a Service«, »Das Meer der Daten«, »Smart Grid – das Internet der Energie« und »Drahtlose Sensornetzwerke« der ersten ÖFIT-Publikation zu Innovationsfeldern aus dem Jahr 2013 rückblickend betrachtet. Für jedes dieser Felder haben FOKUS-Expert:innen die Entwicklung seit 2013 sowie zukünftige Möglichkeiten zusammengefasst.

[6] Ein Innovationsfeld kann einen Trend oder mehrere, sich gegenseitig bedingende Trends beinhalten. Als historisches Beispiel sei hier die Erfindung des Audioformats mp3 genannt. Im Innovationsfeld rundum digitalkomprimierte Medieninhalte waren anschließend Trends wie die Entwicklung und Verbreitung von neuen Dateiformaten, von Tauschbörsen und von kleinen Abspielgeräten beobachtbar.

1. DATENVISUALISIERUNG ERKLÄRT

Begleitend zur inhaltlichen Bearbeitung erfolgte für jedes Innovationsfeld eine Datenauswertung, deren Ergebnis auf verschiedene Weisen visualisiert ist. Um ein besseres Verständnis der Datenvisualisierung zu ermöglichen, sind die wiederholt vorkommenden Grafiktypen hier erläutert. Einmalig vorkommende Grafiktypen sind stattdessen an der Stelle erläutert, an der sie auftreten.

Wissenschaftliche Relevanz

Die Entwicklung der jährlichen Anzahl der wissenschaftlichen Publikationen ist ein wichtiger Indikator für die derzeitige und zukünftige Relevanz eines Trends. Als Datenquelle wurde dabei in der Regel arXiv[7] und im Fall einer zu geringen Ergebnismenge auch Scopus[8] gewählt. arXiv ist ein durch die Cornell University betriebener Dokumentenserver, der es Wissenschafter:innen ermöglicht, ihre Ergebnisse[9] frühzeitig zur Verfügung zu stellen. Beispielsweise wurden 2018 über 140.000 wissenschaftliche Artikel eingereicht.

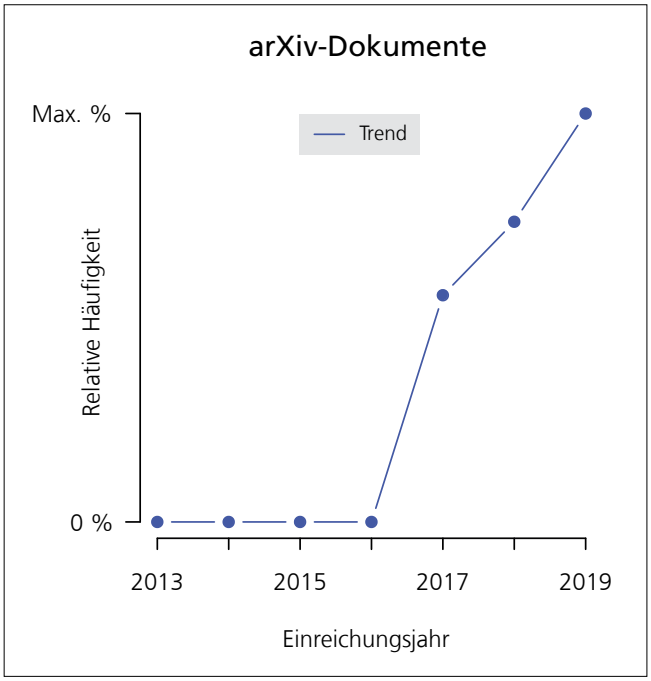


Abb. 1: Beispielhafte Darstellung für den Trendverlauf der wissenschaftlichen Relevanz. Anhand des Graphen lässt sich die Entwicklung des Trends einschätzen, während sich anhand der Obergrenze der y-Achse die bereits erreichte Relevanz ablesen lässt.

Mithilfe passender Suchbegriffe wurden die Artikel zu einem Trend identifiziert. Pro Jahr wurde anschließend die relative Häufigkeit ermittelt, also wie groß der Anteil der Artikel zu diesem Trend an der Gesamtanzahl der Artikel ist. Beispielsweise wurden 2018 507 Artikel zu Big Data veröffentlicht, was einer relativen Häufigkeit von gerundet 0,36 % entspricht. Weil Big Data ein bereits etabliertes und immer noch relevantes Thema ist, kann diese relative Häufigkeit als vergleichsweise hoch angesehen werden. Die relativen Häufigkeiten der Trends innerhalb der Innovationsfelder sind in der Regel deutlich geringer.

Geografische Verortung

Wie intensiv wird in Staaten zu den Trends eines Innovationsfeldes geforscht und wie sind die Staaten vernetzt? Um diese Frage zu beantworten, wurden die Standorte der Forschungseinrichtungen der Autor:innen zu einem Innovationsfeld betrachtet. Eine Publikation wurde einem Staat zugeordnet, wenn zumindest ein:e Autor:in einer Forschungseinrichtung aus diesem Staat angehörte. Datengrundlage waren Scopus-Suchergebnisse für die Publikationsjahre 2014 bis 2019.

Jedes Kreisdiagramm in der Grafik repräsentiert einen Staat, wobei nur Staaten mit einer gewissen Mindestanzahl[10] an Publikation berücksichtigt wurden. Die Fläche des Kreises ist dabei proportional zur jahresübergreifenden Anzahl der Publikationen, während die Flächen der Anteile proportional zu den jahresweisen Publikationshäufigkeiten sind. Weil Veröffentlichungen oftmals nicht sofort durch Literaturdatenbanken wie Scopus erfasst werden und mit absoluten Häufigkeiten gerechnet wurde, ist davon auszugehen, dass die 2019er Anteile in der Regel kleiner ausfallen als sie tatsächlich sind. Sind zwei Staaten durch eine graue Linie verbunden, so bedeutet dies, dass Forscher:innen dieser Staaten gemeinsam publiziert haben.

[7] https://arxiv.org.

[8] https://www.scopus.com.

[9] Dass ein eingereichter Artikel wissenschaftlichen Ansprüchen genügt, wird vor der Veröffentlichung geprüft, allerdings nicht im gleichen Ausmaß wie bei einer wissenschaftlichen Fachzeitschrift. Siehe auch https://arxiv.org/help/submit.

[10] Diese Mindestanzahl ist je nach Innovationsfeld unterschiedlich und schränkt die Anzahl der repräsentierten Staaten ein, wodurch die Übersichtlichkeit der Grafik gewährleistet wird.

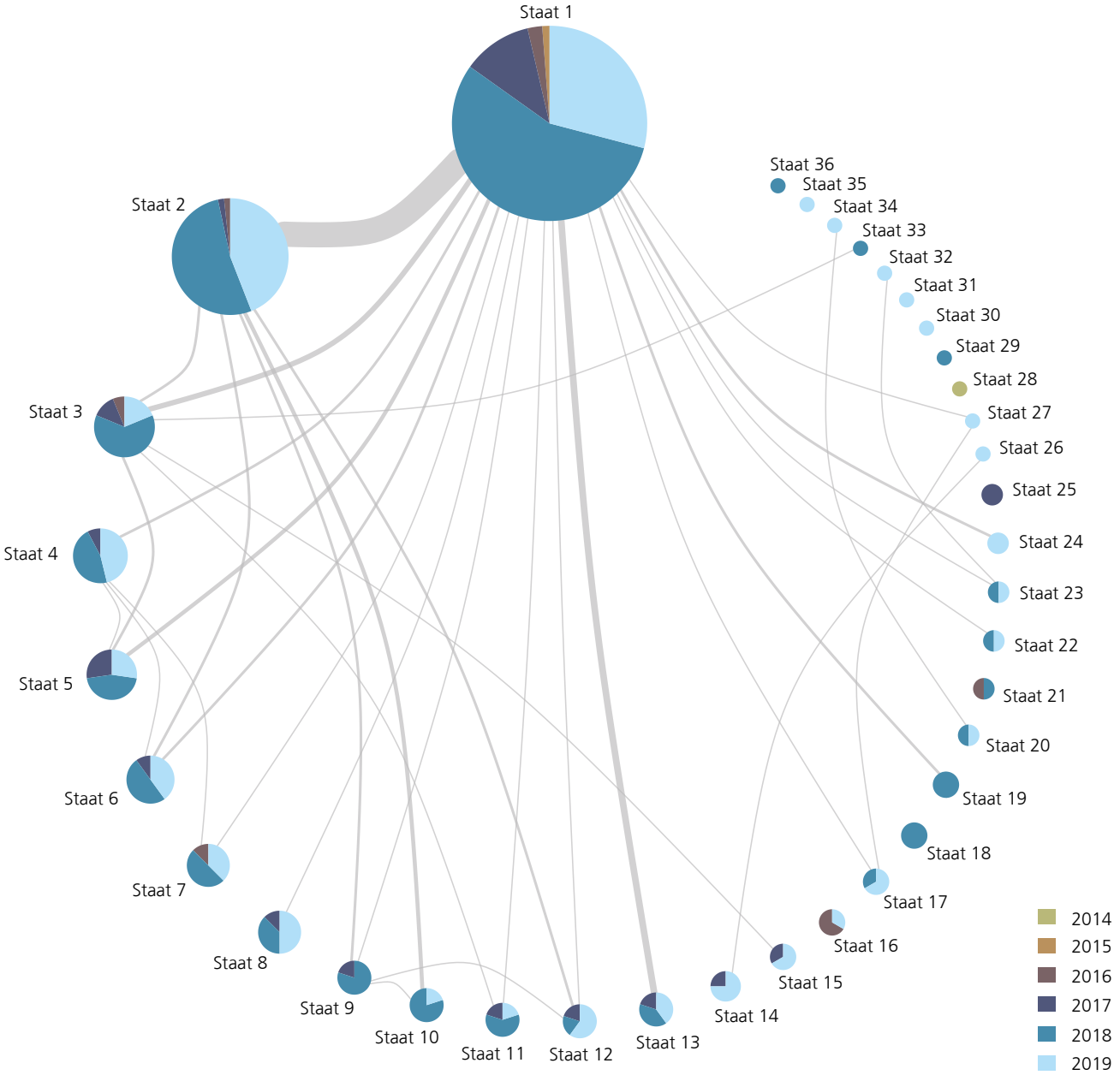


Abb. 2: Beispielhafte Darstellung für die geografische Verteilung der Publikationshäufigkeit. In den Bildbeschreibungen sind die minimalen und maximalen Publikationsanzahlen pro Staat (proportionale Kreisdiagrammfläche) bzw. zwischen Staaten (proportionale Liniendicke) als »Publikationsanzahl« und »Gemeinsame Publikationsanzahl« wiedergegeben, also bspw. »Publikationsanzahl: 1 bis 50. Gemeinsame Publikationsanzahl: 1 bis 11.«

Die Linienbreite ist proportional zur jahresübergreifenden Häufigkeit der Zusammenarbeit – je breiter eine Verbindungslinie ist, desto stärker die Zusammenarbeit. Auf diese Weise wird die internationale Vernetzung zu einem Innovationsfeld deutlich.

Natürlich ist nicht nur die Quantität, sondern auch die Qualität der Forschung von Interesse. Um die Qualität der Forschungsergebnisse zu messen wurde der auch als h-Index bekannte Hirschfaktor verwendet. Dieser wird in der Regel genutzt, um den Einfluss einer Wissenschaftlerin oder eines Wissenschaftlers anhand der Zitationshäufigkeit der Publikationen zu messen. Der Hirschfaktor h ist die größtmögliche Anzahl von Publikationen, die mindestens h-mal zitiert wurden. Deutlich wird dies an einem Beispiel: Angenommen, eine Forscherin hat 5 Publikationen veröffentlicht, die 7-, 3-, 1-, 0- und 0-mal zitiert wurden. Der Hirschfaktor ist dann 2, denn es existieren 2 Publikationen die jeweils mindestens zweimal zitiert wurden, aber keine 3 Publikationen die jeweils mindestens dreimal zitiert wurden. Um die Qualität der Forschungsergebnisse für einen Staat zu messen, wurde diese Berechnungsmethode auf die einem Staat zugeordneten Publikationen angewendet. Das Ergebnis wurde als Balkendiagramm visualisiert. Dabei wurden nur Staaten mit mindestens fünf Publikationen berücksichtigt.

Wahrnehmung in Wissenschaft und sozialen Medien

Um einen Eindruck davon zu vermitteln, welche Themen mit den Innovationsfeldern assoziiert werden und welche Unterschiede diesbezüglich zwischen Wissenschaft und Gesellschaft bestehen, wurde die Verschlagwortung von Beiträgen zum Innovationsfeld grafisch aufbereitet. Dazu wurden englischsprachige Tweets und wissenschaftliche Publikationen zu den Trends des jeweiligen Feldes gesammelt. Anschließend wurden die 50 häufigsten Hashtags von Twitter und die 50 häufigsten Schlüsselwörter aus den wissenschaftlichen Publikationen ermittelt. Die Begriffe sind in Form einer Wordcloud dargestellt. Je populärer ein Begriff ist, desto größer ist er. Anhand der Farbgebung lässt sich erkennen, ob ein Begriff nur als Hashtag, nur als Schlüsselwort oder als Hashtag und Schlüsselwort aufgetreten ist. Falls ein Begriff, der sowohl als Hashtag als auch als Schlüsselwort auftrat, mit einer Raute beginnt, so bedeutet dies, dass die Popularität des Begriffs als Hashtag in den sozialen Medien größer war als die Popularität des Begriffs als Schlüsselwort in der Fachliteratur. Im umgekehrten Fall fehlt die Raute.

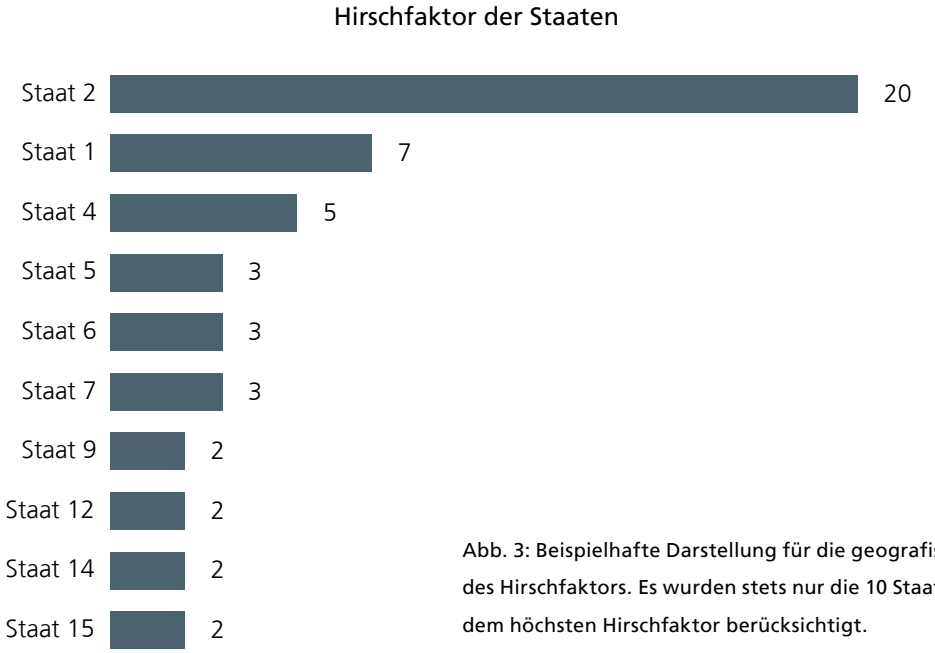
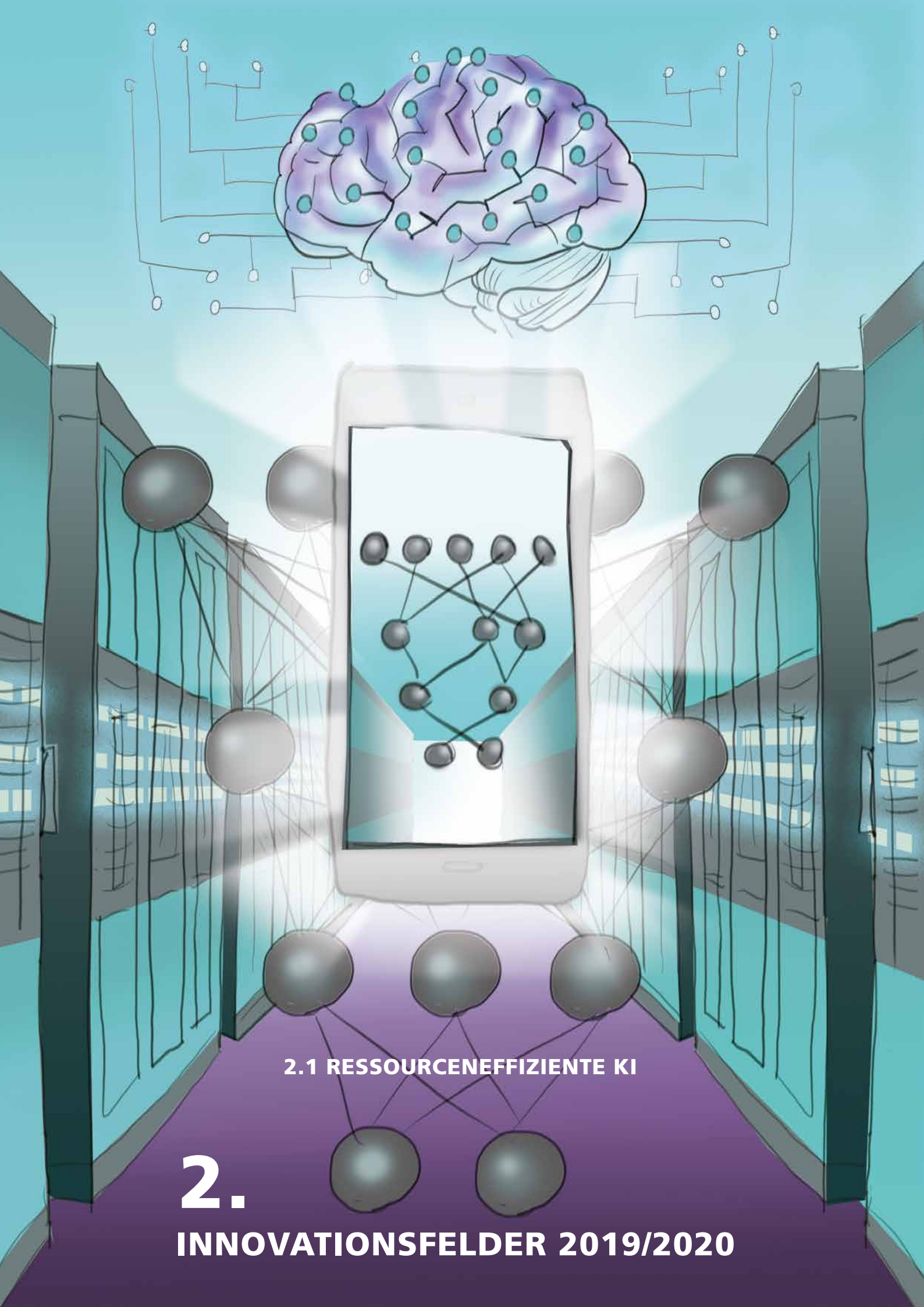


Abb. 3: Beispielhafte Darstellung für die geografische Verteilung des Hirschfaktors. Es wurden stets nur die 10 Staaten mit dem höchsten Hirschfaktor berücksichtigt.

Abb. 4: Beispielhafte Darstellung der Wordcloud zur Wahrnehmung in Wissenschaft und sozialen Medien.



2.1 RESSOURCENEFFIZIENTE KI

2. INNOVATIONSFELDER 2019/2020

KI-Anwendungen sind heutzutage vorrangig rechenintensive Programme, die auf Hochleistungsrechnern ausgeführt werden. Training, Datenverarbeitung und weitere Dienste sind dort zentralisiert. Mobilgeräte, die KI-Anwendungen nutzen, beschränken sich bislang auf den Datenaustausch mit den KI-Rechenzentren der Dienstanbieter. In diesem Anwendungsgebiet zeichnet sich ein Dezentralisierungstrend ab: Es werden zunehmend KI-Technologien, Trainingsmethoden und spezielle Hardware entwickelt, durch die KI-Anwendungen unabhängig von großen KI-Rechenzentren operieren und stattdessen beispielsweise auf mobilen Geräten ausgeführt werden können. Durch ressourceneffiziente KI lassen sich neue Anwendungsmöglichkeiten von KI realisieren. Diese beeinflussen, wie KI in Zukunft entwickelt, vertrieben und genutzt wird. Die Auswirkungen ressourceneffizienter KI können dabei aus gesellschaftlicher, wirtschaftlicher und ökologischer Perspektive betrachtet werden.

Anwendungsmöglichkeiten und Ziele

Ressourceneffiziente KI-Anwendungen können auf Smartphones, IoT-Geräten und Edge Devices, bei autonomen Fahrzeugen sowie im Bereich Augmented & Virtual Reality zum Einsatz kommen. Beispiele für Aufgaben, die durch ressourceneffiziente KI dezentralisiert werden, sind Bild- und Spracherkennung, Mustererkennung in Nutzerdaten, Bild- und Audiobearbeitung (siehe Kapitel 2.2) sowie intelligentes und adaptives Routing von Netzwerkverkehr. Durch die Dezentralisierung werden einige KI-Anwendungen schneller, da Latenzzeiten des Netzwerks wegfallen – für bestimmte Anwendungen, bspw. in autonomen Fahrzeugen, ist dies sogar eine notwendige Voraussetzung. Jedoch erfordert das Aktualisieren dezentraler KI-Anwendungen (etwa für die Fehlerbehebung) wieder Softwareupdates in herkömmlicher Manier, statt nur die Software des KI-Rechenzentrums zu ersetzen. Hierbei entstehen neue Schwachstellen und Angriffspotenziale (siehe Kapitel 2.4). Ein Nebeneffekt der Dezentralisierung ist, dass KI-Anwendungen datenschutzkonformer werden können. Nutzerdaten müssen zur Verarbeitung nicht mehr an KI-Rechenzentren versendet werden, sondern verbleiben auf den Nutzergeräten. Zukünftig können Daten sogar teilweise in verschlüsseltem Zustand verarbeitet werden, um ansteigenden Datensicherheitsanforderungen gerecht zu werden.

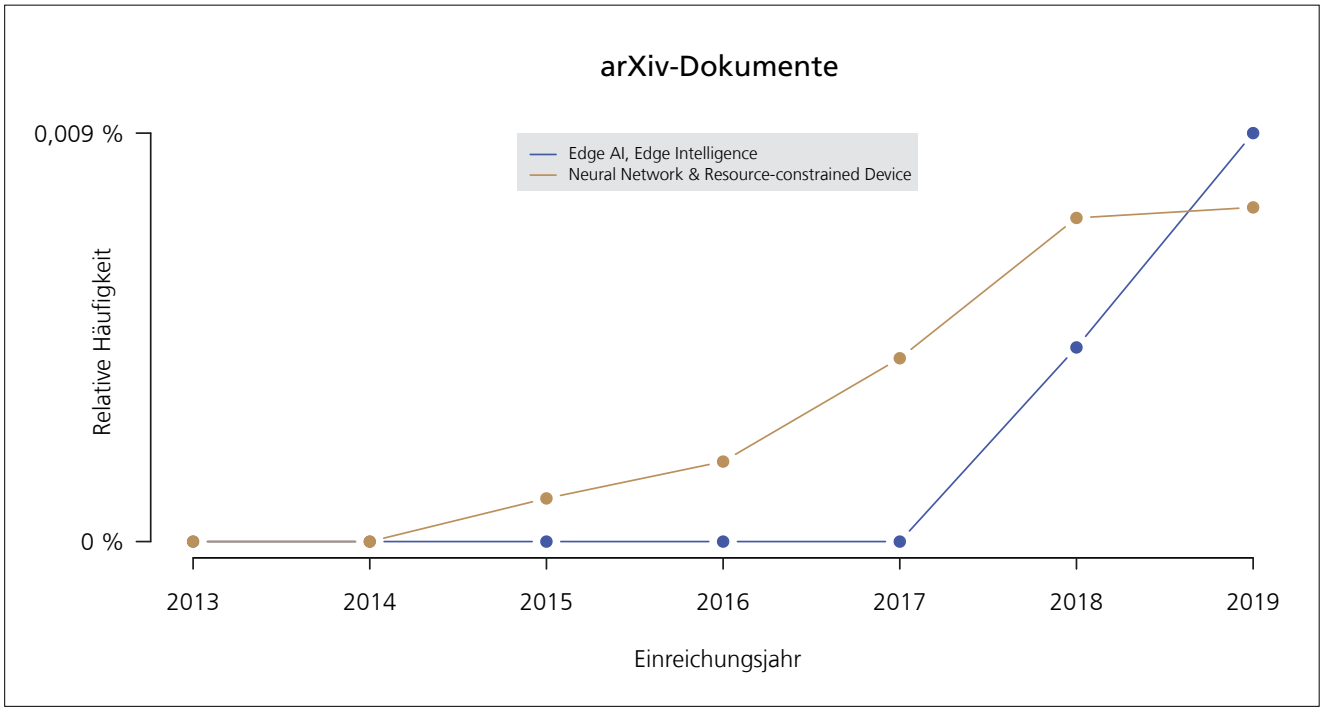


Abb. 5: Trendverlauf für Anwendungsmöglichkeiten und Ziele ressourceneffizienter KI nach ihrem Vorkommen im Titel oder Abstract. Erhebung vom 6.11.2019.

Software

Ressourceneffiziente KI wird durch die Weiterentwicklung der zugrunde liegenden mathematischen Modelle, Trainingsalgorithmen und Auswertungsmechanismen ermöglicht. Das Forschungsgebiet Machine Learning[11] und insbesondere dessen Teilbereich Deep Learning befindet sich seit einigen Jahren im Aufschwung. Dabei werden neue Netzarchitekturen für künstliche neuronale Netze[12] entworfen und Trainingsalgorithmen verbessert, die maßgeblich zu Güte und Effizienz der daraus resultierenden KI-Anwendungen beitragen.

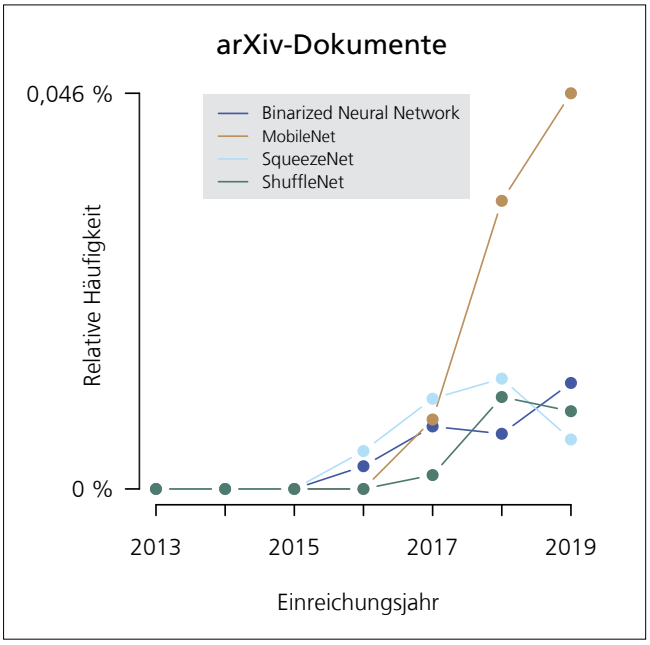


Abb. 6: Trendverlauf für Typen ressourceneffizienter künstlicher neuronaler Netze nach ihrem Vorkommen im Titel oder Abstract. Erhebung vom 6.11.2019.

Forscher:innen drehen bei der Weiterentwicklung von KI an mehreren Stellschrauben gleichzeitig: Der Anzahl der Trainingsparameter, der Größe und Beschaffenheit der neuronalen Netze und den Auswertungsmechanismen. Außerdem werden neue Techniken entworfen, um einzelne Rechenoperationen effizienter zu gestalten und architekturelle Vorteile zu nutzen. Beispiele dafür sind der Einsatz von binären Operatoren statt ganzzahliger, der Einsatz von sogenannten Convolutions (Faltungen) in der Netzarchitektur und die Verkleinerung der Netzparameter durch neu entwickelte Kompressionsmethoden.

Die Methode des Neural Network Pruning bietet zudem einen Ansatz, bereits vollständig trainierte neuronale Netzwerke im Nachhinein zu verkleinern. Oft können Teile von neuronalen Netzwerken entfernt werden, ohne dass das Netzwerk dadurch seine Leistungsfähigkeit einbüßt. Es soll dabei im bildlichen Sinn aus dem gegebenen Netzwerk ein kleineres, aber bezüglich der Anwendung annähernd gleichwertiges Netzwerk herausgemeißelt werden. Durch diese Methode lässt sich die Größe neuronaler Netzwerke um einen Faktor von 10 bis 100 reduzieren.[13] Neural Network Pruning könnte sich im Vergleich zu den zuvor beschriebenen Ansätzen als bedeutender für die Verbreitung ressourceneffizienter KI erweisen. Zwar muss das Training zentralisiert und mit hohem Ressourcenaufwand ausgeführt werden, das trainierte KI-System kann aber dann lokal eingesetzt werden und ist damit in der Anwendungsphase ressourceneffizienter.

Durch die geringere Größe und die effizienteren Ausführungsmechanismen dieser neuen neuronalen Netze können sie einfacher und kostengünstiger über das Internet übertragen und auch auf leistungsschwächeren Geräten ausgeführt werden.

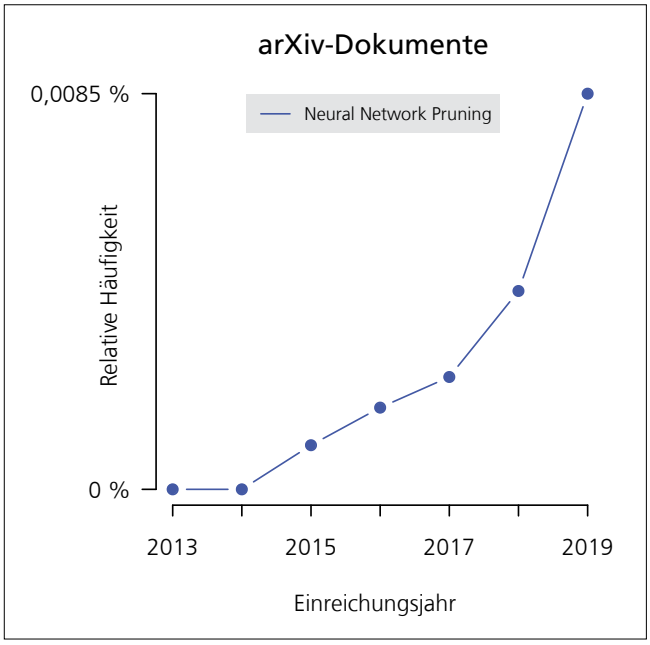


Abb. 7: Trendverlauf für Neural Network Pruning nach Vorkommen im Titel oder Abstract. Erhebung vom 27.11.2019.

[11] Wesentliches Teilgebiet der KI.

[12] Durch das menschliche Gehirn inspirierte Machine-Learning-Modellarchitekturen. Siehe auch: Grosch, D. (2019).

[13] Frankle, J.; Carbin, M. (2019).

Hardware

Ressourceneffiziente KI wird nicht nur durch Fortschritte in der Software ermöglicht. Die ressourcenintensiven Rechenoperationen neuronaler Netze werden auch durch spezielle Hardware wie KI-Prozessoren effizienter ausgeführt. Insbesondere für batteriebetriebene Mobilgeräte wie Smartphones sind Hardware-Effizienzgewinne bei KI-Anwendungen wichtig. Spezielle Hardware für KI-Anwendungen hat ein neues Marktsegment eingeläutet. Immer mehr Chiphersteller produzieren zusätzlich zu herkömmlichen Prozessoren (CPUs und bisher dominierende Hardwarebeschleuniger wie etwa GPUs und TPUs) sogenannte neuromorphe Prozessoren (NPUs), deren elektronische Strukturen auf die Architektur von neuronalen Netzen zugeschnitten sind. Durch die Kopplung mit in Hardware »gegossenen« Auswertungsmechanismen können KI-Algorithmen um ein Vielfaches schneller oder auch energiesparender ausgeführt werden. NPUs sind eine neue Technologie, in der sich noch kein dominantes Design durchgesetzt hat. Es gibt zwei Varianten von NPUs. Erstere implementieren bewährte und erfolgreiche Trainingsalgorithmen als Hardwarekomponenten und beschleunigen den aufwändigen Trainingsprozess dadurch signifikant. Bei Letzteren werden bereits trainierte neuronale Netze temporär auf die neuromorphe Hardware »geladen« und so von der Hardware beschleunigt ausgeführt.

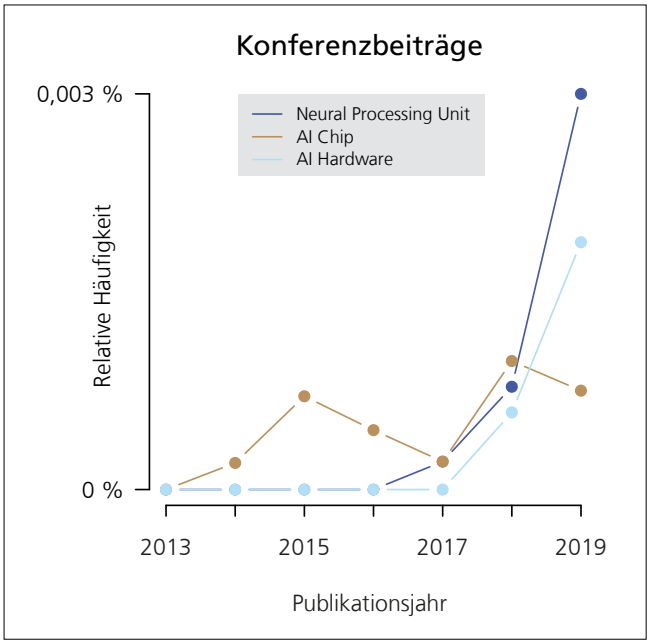


Abb. 8: Trendverlauf für Hardware für ressourceneffiziente KI nach Vorkommen im Titel oder Abstract. Datenquelle: Scopus. Erhebung vom 6.11.2019.

Auswirkungen, Chancen und Risiken

Ressourceneffiziente KI ist ein entscheidender Schritt hin zur Allgegenwärtigkeit von KI-Anwendungen. Diese Entwicklung kann langfristig zum flächendeckenden Einsatz lokal ausgeführter KI-Anwendungen und neuromorpher Hardwarechips auf Mobilgeräten führen. Die durch ressourceneffiziente KI erreichte Lokalität von KI-Anwendungen bietet Chancen für einen besseren Datenschutz und mehr Nutzerkontrolle. KI-Anwendungen können dabei sogar auf Betriebssystemebene eingebunden werden und somit eine noch passgenauere Individualisierung ermöglichen. Dabei ist ressourceneffiziente KI auch aus einer ökologischen Perspektive interessant: Die Verringerung der Ressourcenanforderungen für Training und Einsatz von KI könnte sich positiv auf die Klimabilanz niederschlagen, allerdings ist hier auch ein Rebound-Effekt[14] denkbar.

Die Anforderungen an die für das Trainieren und Ausführen performanter KI-Anwendungen notwendige Rechenleistung werden durch ressourceneffiziente KI geringer. Wenn keine Rechenzentren mehr notwendig sind oder zumindest kostengünstiger und breiter verfügbare Server ohne GPUs/TPUs nutzbar werden, könnte dies zu einer neuen Welle von KI-Anwendungen führen, die auch von kleineren Entwicklungsteams bzw. sogar von Privatpersonen programmiert werden. Somit könnte ressourceneffiziente KI potenziell zu mehr Chancengerechtigkeit bei KI beitragen.

Handlungsempfehlungen

Trainingsdaten in Bezug auf Datenschutz regulieren. Ressourceneffiziente KI ist eine Chance aber keine Garantie für verbesserten Datenschutz. Daher gilt, wie auch bei KI im Allgemeinen, dass die Gewinnung und Nutzung (insbesondere sensibler) nutzergenerierter Trainingsdaten aus einer datenschutzrechtlichen Perspektive bewertet und reguliert werden müssen.

Forschung zu ressourceneffizienten neuronalen Netzen fördern. Die Förderung der Grundlagenforschung zu künstlichen neuronalen Netzen und deren Überführung zu Anwendungen ist von zentraler Bedeutung für den (wissenschaftlichen) Fortschritt in diesem Innovationsfeld. Insbesondere unkonventionelle und experimentelle Ansätze könnten zu Fortschritten bei der Ressourceneffizienz führen.

[14] Beim Rebound-Effekt kommt es zu einem insgesamt erhöhten Ressourcenverbrauch, weil ein neuerlich ressourceneffizienteres Gut nun deutlich häufiger genutzt wird. Siehe auch https://www.umweltbundesamt.de/themen/abfall-ressourcen/oekonomische-rechtliche-aspekte-der/rebound-effekte.

Geografische Verortung

Im Innovationsfeld »Ressourceneffiziente KI« sind, wie beim allgemeinen Trend der KI, die Volksrepublik China und die USA die führenden Forschungsnationen. Es lässt sich eine hochgradige Kooperation zwischen den beiden Nationen erkennen. Die Forschungskooperation bei ressourceneffizienter KI ist aber auch sonst sehr umfangreich. Im Vergleich zu den anderen Innovationsfeldern liegt Deutschland bezüglich der Publikationshäufigkeit etwas weiter zurück.

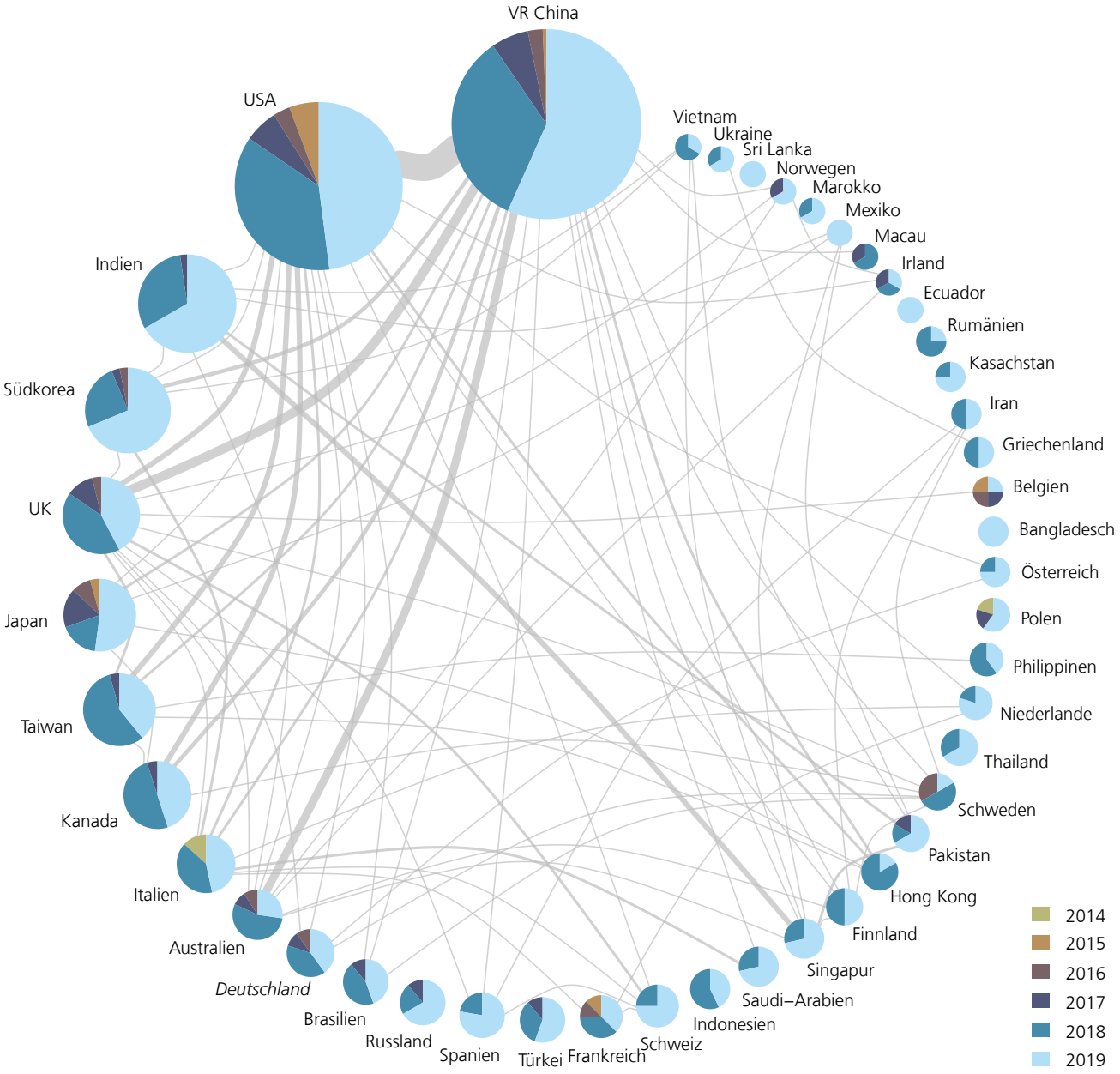


Abb. 9: Geografische Verteilung der Publikationshäufigkeit für ressourceneffiziente KI. Publikationsanzahl: 3 bis 157. Gemeinsame Publikationsanzahl: 1 bis 20. Erhebung vom 08.11.2019.

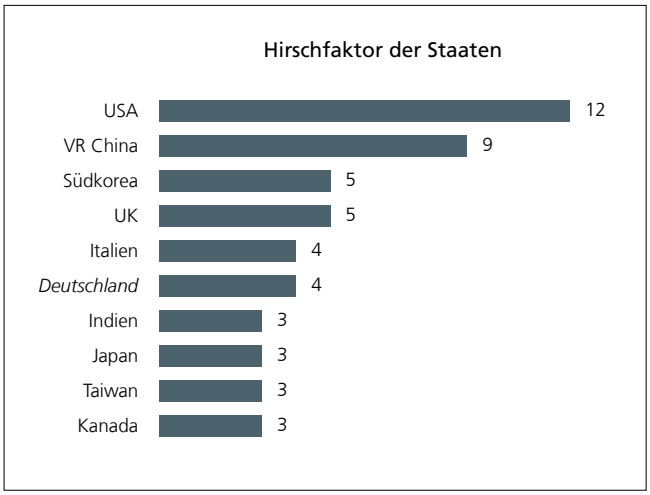


Abb. 10: Geografische Verteilung des Hirschfaktors für ressourceneffiziente KI. Erhebung vom 08.11.2019.

Wahrnehmung in Wissenschaft und sozialen Medien

Die Namen bekannter Hersteller von Hardware und Software für KI fallen in den sozialen Medien besonders häufig, was sich möglicherweise als Diskussion über ressourceneffiziente KI-Produkte dieser Hersteller interpretieren lässt. Die wissenschaftlichen Publikationen befassen sich hingegen stark mit den unterschiedlichen Architekturen neuronaler Netze. Anwendungsmöglichkeiten wie Edge Computing und das Internet der Dinge werden sowohl in der Wissenschaft als auch den sozialen Medien wahrgenommen.

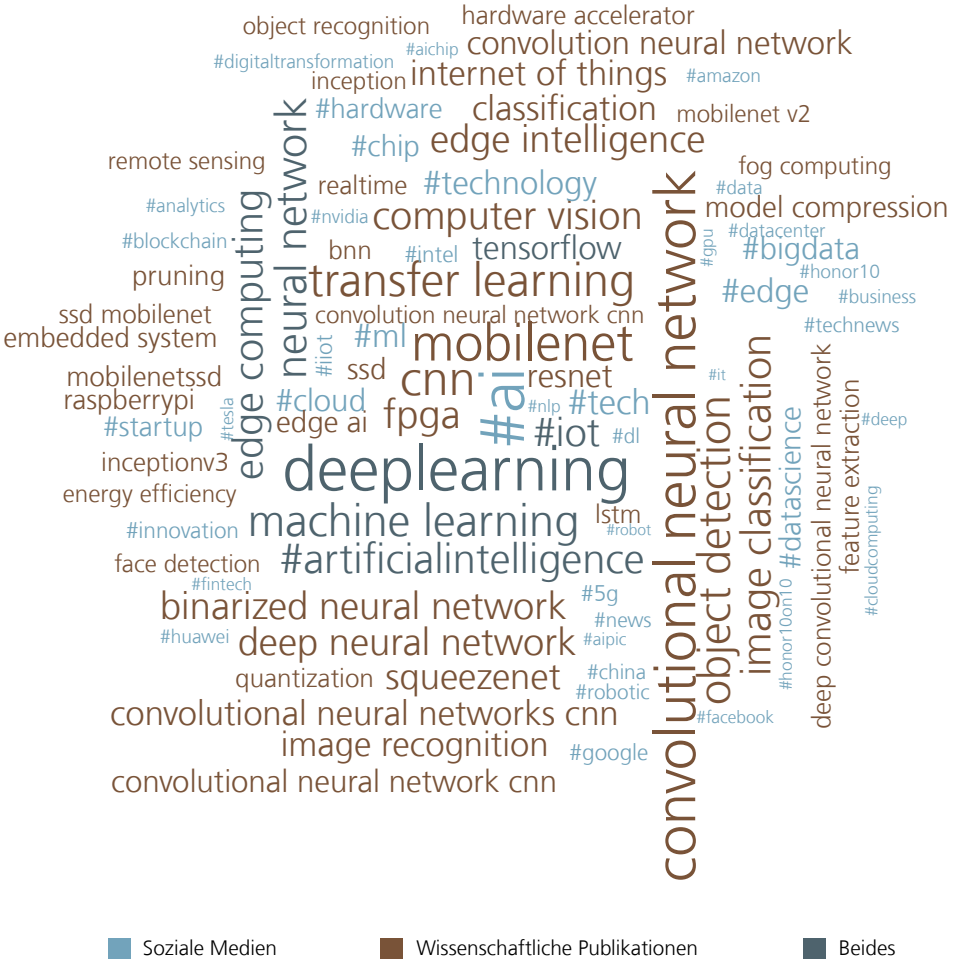
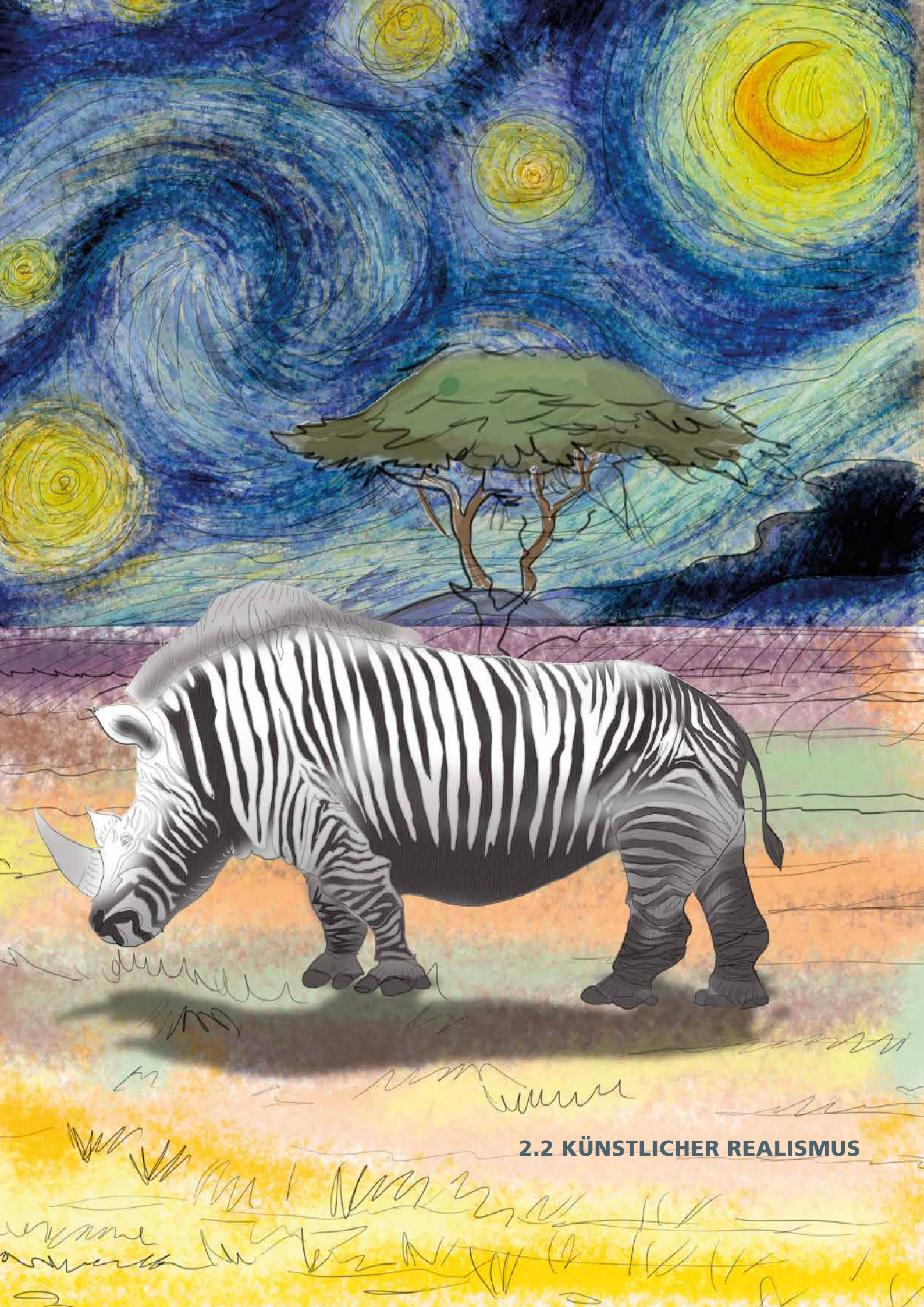


Abb. 11: Wordcloud für das Innovationsfeld ressourceneffiziente KI. Die populärsten Hashtags und Schlüsselwörter für die Jahre 2018 und 2019. Erhebung vom 15.11.2019.



2.2 KÜNSTLICHER REALISMUS

Authentisch wirkende Bild-, Video- und Audioaufnahmen sowie Texte mit teilweise oder vollständig künstlichem Ursprung sind nicht neu. Bei der Erstellung kamen dabei in der Vergangenheit beispielsweise Computer Generated Imagery[15], Bildbearbeitungsprogramme und Stimmimitator:innen zum Einsatz. Vor wenigen Jahren erfolgte ein auf neuartigen Architekturen künstlicher neuronaler Netze beruhender technologischer Durchbruch. Während die Ergebnisse rasch überzeugender werden, ist das Alleinstellungsmerkmal gegenüber etablierten Techniken ein anderes: Die notwendigen Voraussetzungen bezüglich Aufwand und Fachwissen für die massenhafte Erzeugung dieser Inhalte werden geringer.

Style Transfer

In der Informatik versteht man unter Style Transfer die Erzeugung von Texten sowie Audio-, Bild- oder Videomaterial mit zwei verschiedenen Eingaben. Dabei soll der Inhalt der einen Eingabe und der Stil der anderen Eingabe übernommen werden. Beispielhaft lässt sich dies am Style Transfer für Bilder erklären. Bei der ersten Eingabe handelt es sich bspw. um eine Fotoaufnahme des Empire State Buildings, während die zweite Eingabe ein Gemälde Vincent van Goghs ist. Die gewünschte Ausgabe ist ein Bild, das das Empire State Building so zeigt, als wäre es von van Gogh gemalt worden. Der Anspruch auf Realismus besteht hierbei in der möglichst getreuen Nachahmung des künstlerischen Stils Vincent van Goghs. Image Style Transfer wird zum Beispiel durch Neural Style Transfer ermöglicht. Hierbei handelt es sich um einen Sammelbegriff für Methoden, die auf Objekterkennung spezialisierte künstliche neuronale Netze verwenden. Mithilfe von Objekterkennung werden Inhalt und Stil der Eingaben getrennt, um dann den Stil zu übertragen. Bilder sind in der Fachliteratur besonders häufig Gegenstand der Forschung zu Style Transfer, allerdings ist Style Transfer auch für Video- und Audioaufnahmen, Text[16] sowie Typografie[17] möglich. Während künstliche neuronale Netze ein sehr populärer Lösungsansatz für Style Transfer sind, wurden auch ohne neuronale Netze bereits überzeugende Ergebnisse erreicht.[18]

[15] Bildsynthese, die beispielsweise in der Filmindustrie zum Einsatz kommt.

[16] Prabhumoye, S.; Tsvetkov, Y.; Salakhutdinov, R.; Black, A. W. (2018).

[17] Wenige Zeichen dienen als Stilvorlage, anhand derer dann ein Zeichensatz im gleichen Stil erzeugt wird, wodurch eine komplette Schriftart entsteht. Siehe auch Yang, S.; Liu, J.; Lian, Z.; Guo, Z. (2016).

[18] Siehe auch Yang, S.; Liu, J.; Lian, Z.; Guo, Z. (2016) und Jamriška, O.; Sochorová, Š.; Texler, O.; Lukáč, M.; Fišer, J.; Lu, J.; Shechtman, E.; Sýkora, D. (2019).

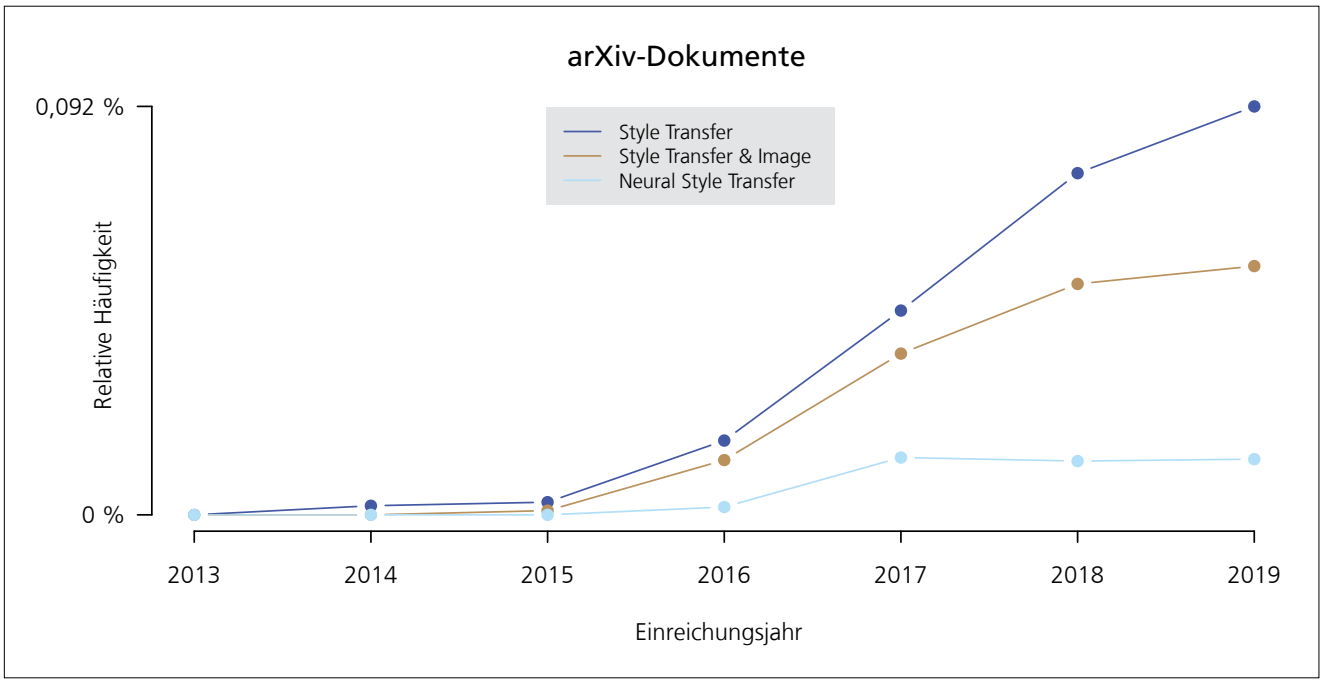


Abb. 12: Trendverlauf für Style Transfer nach Vorkommen im Titel oder Abstract. Erhebung vom 14.10.2019.

Übersetzungsprobleme

Viele Probleme, bei denen Audio-, Bild- oder Videomaterial in realistischer Weise verändert werden soll, lassen sich als Übersetzungsprobleme auffassen. Das Prinzip des Übersetzungsproblems lässt sich am Beispiel mathematischer Funktionsvorschriften erklären. Mathematische Funktionen ordnen den Elementen einer Definitionsmenge Elemente einer Wertemenge zu und die Funktionsvorschrift gibt an, wie das geschieht. Bei Übersetzungsproblemen sind zwei Mengen von Beispielen gegeben und anhand dieser Beispiele soll ein Computer nun eine Zuordnungsvorschrift von der ersten zur zweiten Menge lernen, die dann auf neue Eingaben anwendbar ist. Image-to-Image Translation ist eine Klasse von Übersetzungsproblemen, der in den letzten Jahren verstärkte Aufmerksamkeit zuteilwurde. Beispielhaft lässt sich Image-to-Image Translation anhand von Landschaftsfotos erklären: Angenommen, für Fotos von Landschaften während des Sommers sollen plausible Darstellungen der gleichen Landschaften während des Winters erzeugt werden. Diese Aufgabe lässt sich als Image-to-Image Translation formulieren: Anhand von Beispielbildern von Landschaften während des Sommers und Winters soll eine Zuordnungsvorschrift erlernt werden, um die zu erwartenden Auswirkungen des Winters auf das visuelle Erscheinungsbild der Sommerlandschaften zu übertragen. Dieses Beispiel zeigt, dass sich Style Transfer als Image-to-Image Translation Problem formulieren lässt. Das Erhöhen der Auflösung bei Bild- und Videoaufnahmen (Superresolution), die Beseitigung von Artefakten (z. B. Rauschen) bei Audio-, Bild-, und Videoaufnahmen sowie die Erzeugung synthetischer Aufnahmen sind weitere Aufgaben, für die eine Formulierung als Übersetzungsproblem möglich ist.

Die technologischen Fortschritte bei Image-to-Image Translation beruhen auf Encoder-Decoder Networks enthaltende Generative Adversarial Networks. Bei Generative Adversarial Networks übernehmen Teile der Architektur die Aufgabe, synthetische Bilder zu erzeugen (Generators), während andere Teile die erzeugten Bilder auf ihre Authentizität prüfen (Discriminators). Durch eine Art Wettlauf optimieren sich Generators und Discriminators so gegenseitig. Die eingesetzten neuronalen Netze komprimieren dabei Bilddaten und extrahieren so die wesentlichen Bildinhalte (Encoder), um anschließend anhand dieser Bildinhalte wieder vollständige Bilder durch Dekompression zu konstruieren (Decoder). Die Architekturen der Generative Adversarial Networks unterscheiden sich dabei durch Anzahl, Aufbau und Anordnung von Generators und Discriminators. Besonders populär sind etwa Pix2Pix[19] und CycleGAN[20]. Während für Pix2Pix paarweise Trainingsdaten (etwa pro Land-

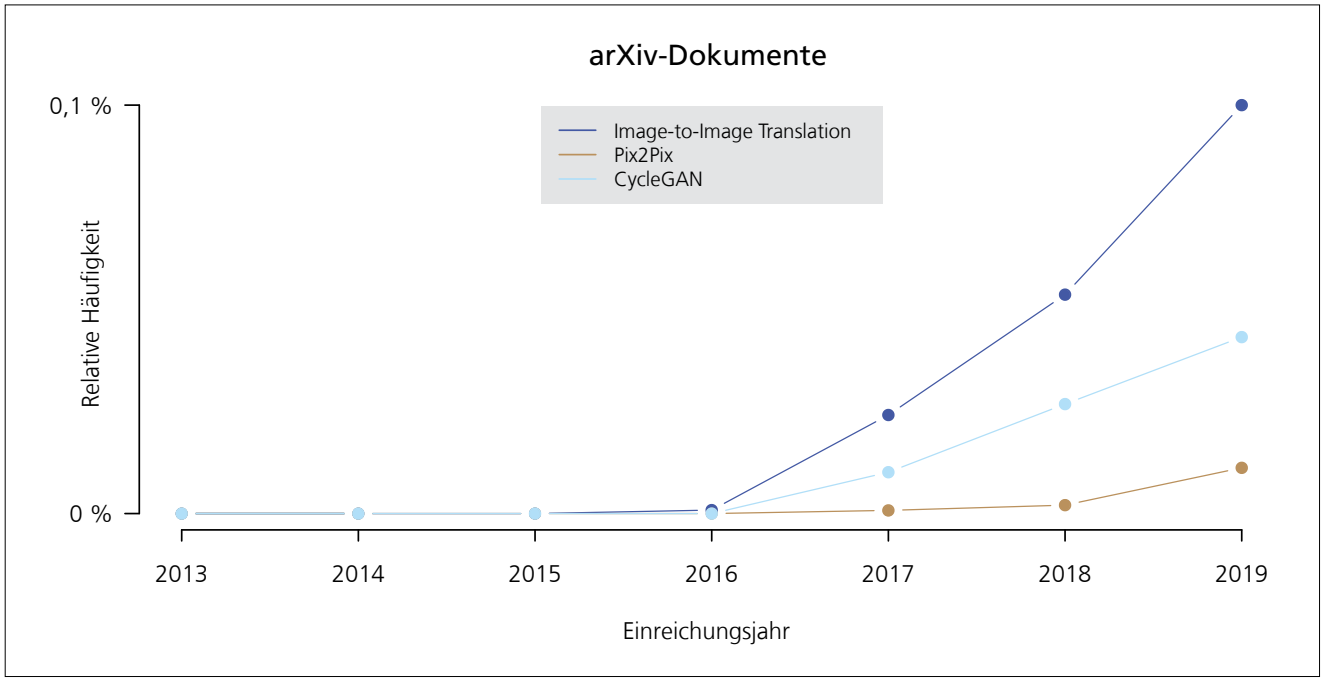


Abb. 13: Trendverlauf für Übersetzungsprobleme nach Vorkommen im Titel oder Abstract. Erhebung vom 15.10.2019.

[19] Verschiedene Implementierungen und Link zur Publikation unter https://phillipi.github.io/pix2pix/.

[20] Verschiedene Implementierungen und Link zur Publikation unter https://junyanz.github.io/CycleGAN/.

schaft und Blickwinkel ein Bild während des Sommers und ein Bild während des Winters) benötigt werden, entfällt diese Anforderung bei CycleGAN, was die Entwicklung erleichtert.

Deepfakes

Wenn Medien über manipulierte Audio-, Bild- oder Videoaufnahmen sowie Texte berichten, wird in den letzten Jahren oft das aus Deep Learning und Fake zusammengesetzte Kofferwort Deepfake verwendet. Dabei existiert bisher keine eindeutige Begriffsdefinition. So wird Deepfake etwa sowohl für die manipulierten Aufnahmen als auch für die dahinterstehende Technik verwendet. Der kleinste gemeinsame Nenner hierbei ist, dass es sich um mittels künstlicher Intelligenz erstellte Audio-, Bild-, Text- oder Videofälschungen handelt.

Zu den diskutierten und teilweise bereits beobachteten Einsatzmöglichkeiten von Deepfakes gehören

- Museumsattraktionen, beispielsweise wird im Dalí Museum in Florida ein Deepfake des Künstlers als interaktives Ausstellungsstück genutzt,[21]
- Betrug durch Social Engineering, beispielsweise ermöglicht die Deepfake-Technologie, die Stimmen von Unternehmensführer:innen zu imitieren,[22]
- Wahlmanipulation, etwa durch die Verbreitung rufschädigender Deepfakes der Kandidat:innen[23],
- Pornografie, bei der Gesichter in Videos ausgetauscht werden,[24] und
- automatisierte und dadurch massenhafte Generierung von Fake-News-Artikeln.[25]

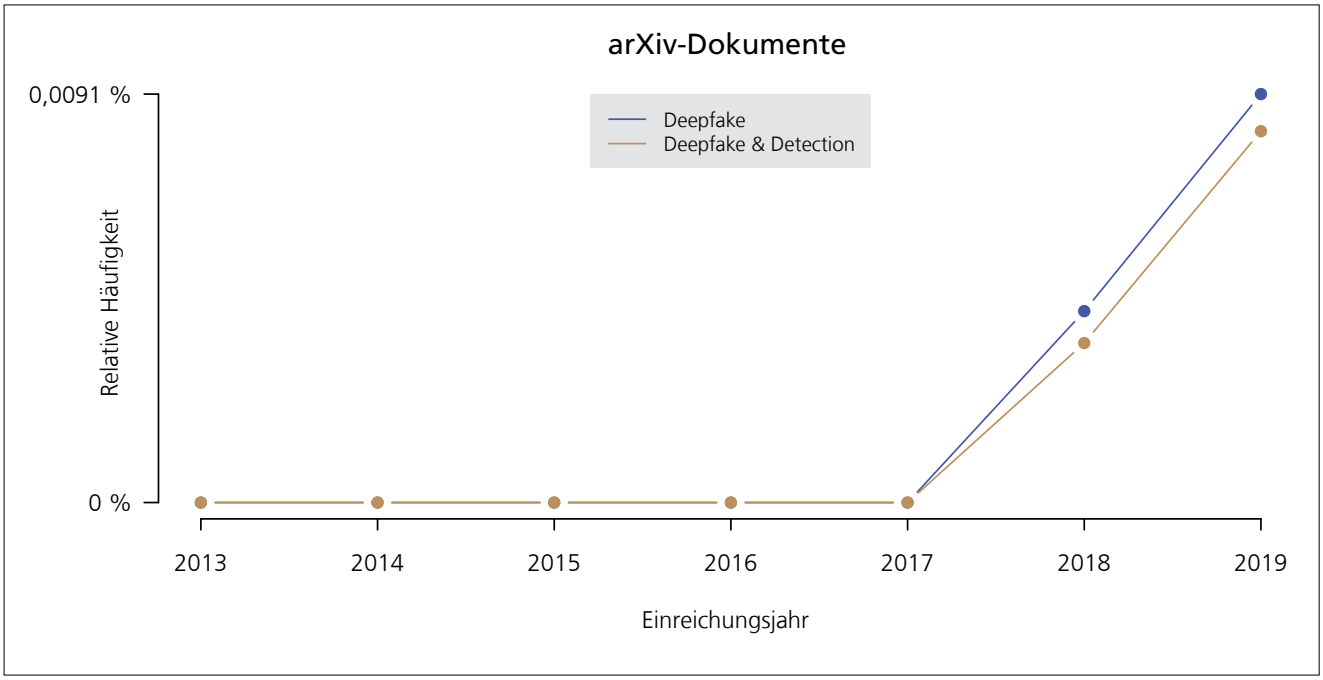


Abb. 14: Trendverlauf für Deepfakes nach Vorkommen im Titel oder Abstract. Erhebung vom 15.10.2019.

Aufgrund des Schadenspotenzials existiert ein Interesse an der Entwicklung von Methoden, mithilfe derer feststellbar ist, ob es sich bei Medieninhalten um Deepfakes handelt. Solche Methoden funktionieren jedoch oftmals nur kurze Zeit und helfen dabei, Schwächen bei Deepfakes zu identifizieren und zu eliminieren.[26] Doch selbst wenn sich Deepfakes als Fälschungen entarnen lassen, taugen sie trotzdem dazu, Personen oder Personengruppen Schaden zuzufügen. Denkbar ist etwa der Einsatz offensichtlicher Fälschungen als Mobbingwerkzeug.

[21] Dalí Museum (2019).

[22] BBC (2019).

[23] Shao, G. (2019).

[24] Kühl, E. (2018).

[25] OpenAI (2019).

[26] Vincent, J. (2019).

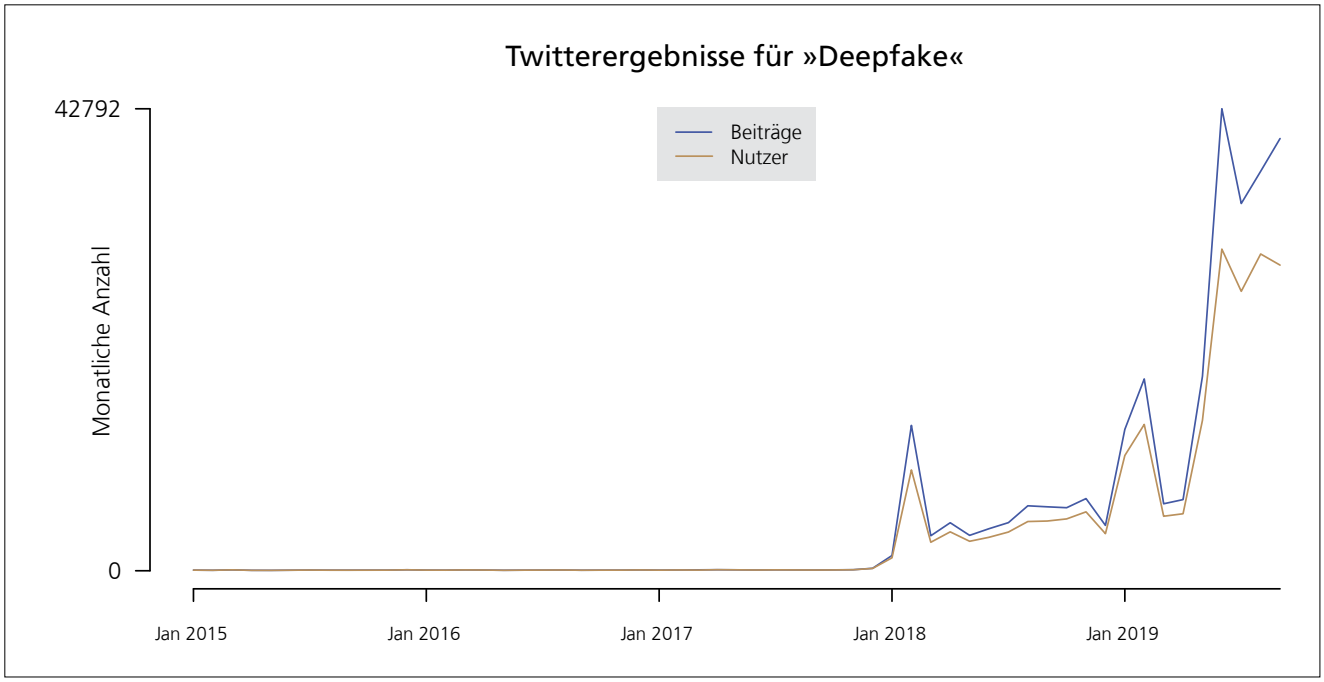


Abb. 15: Trendverlauf für Deepfakes nach Vorkommen im Kurztext. Erhebung vom 17.10.2019

Weitere Anwendungen

Das Entfernen von Bildrauschen, die Korrektur verschwommener Bilder oder das Erzielen einer höheren Bildauflösung sind Aufgaben, die zum Beispiel für die medizinische Bildgebung relevant sind. Image-to-Image Translation kann für solche Aufgaben eingesetzt werden und könnte daher für Fortschritte in der Diagnostik sorgen, weshalb spezifische Lösungsansätze erforscht werden.[27] In der Astronomie wurde Image-to-Image Translation erprobt, um Bildrauschen zu beseitigen und so präzisere Informationen zu Himmelskörpern zu erhalten.[28]

Plausible synthetische Bildaufnahmen lassen sich auch (massenhaft) mittels zufälligen Rauschens erzeugen. Dadurch könnten etwa große Trainingsdatenbestände mit hohem Anonymisierungsgrad[29] für Machine-Learning-Systeme erstellt werden.[30]

Weiterhin könnte in einer Vielzahl von Branchen die Erstellung von Entwürfen beschleunigt werden, indem anhand von Skizzen variantenreiche fotorealistische Abbildungen erstellt werden.[31] Dies könnte beispielsweise beim Design von Kleidung zum Einsatz kommen.[32]

Außerdem wird Image Translation als Möglichkeit zur Erstellung echtzeitfähiger fotorealistischer Avatare erforscht, um soziale Interaktionen innerhalb virtueller Realitäten zu verbessern.[33]

Stilgetreue Textgenerierung könnte bei Schreibassistenten, Chatbots und Übersetzungsprogrammen eingesetzt werden.

[27] Armanious, K.; Jiang, C.; Fischer, M.; Küstner, T.; Nikolaou, K.; Gatidis, S.; Yang, B. (2019).

[28] Shirasaki, M.; Yoshida, N.; Ikeda, S. (2019).

[29] Die Erzeugung synthetischer Daten aus echten Daten ist eine Methode, um anonymisierte Daten zu erhalten. Bei sensitiven Bilddaten ist eine Anonymisierung mitunter notwendig. Zu Anonymisierung siehe auch Gumz, J. D.; Weber, M.; Welzel, C. (2019).

[30] Ein Verwendungszweck ist bspw. Machine Learning in der medizinischen Diagnostik, siehe auch Kaji, S.; Kida, S. (2019).

[31] Zhu, J.; Zhang, R.; Pathak, D.; Darrell, T.; Efros, A. A.; Wang, O.; Shechtman, E. (2018).

[32] Date, P.; Ganesan, A.; Oates, T. (2017).

[33] Wei, S.; Saragih, J.; Simon, T.; Harley, A. W.; Lombardi, S.; Perdoch, M.; Hypes, A.; Wang, D.; Badino, H.; Sheikh, Y. (2019).

Auswirkungen, Chancen und Risiken

In den letzten Jahren sind völlig neue Möglichkeiten zur Erzeugung künstlicher, realistisch wirkender Medieninhalte geschaffen worden. Die niedrigeren Hürden bezüglich Fachwissen und Aufwand führen zu einer Demokratisierung der Erstellung solcher Inhalte. Dadurch können Aufgaben, für die bisher Fachkräfte zuständig waren, nun von Personen mit einem anderen beruflichen Hintergrund bearbeitet und teilweise bis vollständig durch Maschinen erledigt werden. Daher sind auch Änderungen auf dem Arbeitsmarkt wahrscheinlich. Es ist zu erwarten, dass die Verbreitung künstlich realistischer Inhalte weiter zunehmen und die Technik für immer neue Ziele eingesetzt werden wird. Die Eignung von Audio-, Bild-, und Videoaufnahmen als Tatsachenbeleg wird damit jedoch infrage gestellt. Dies könnte Auswirkungen auf die Meinungsbildung innerhalb der Bevölkerung haben. Denkbar sind etwa ein generelles Misstrauen oder ein gesundes Ausmaß an Skepsis gegenüber Medieninhalten. Weil es schwerer ist, den eigenen Sinnen zu trauen, könnten technische Möglichkeiten zur Sicherung der Authentizität von Aufnahmen populärer werden. Diskutiert wird beispielsweise der Einsatz der Blockchain-Technologie, um die Rückverfolgung von Inhalten zu ihrer Quelle zu ermöglichen.[34]

[34] Hasan, H. R.; Salah, K. (2019).

Handlungsempfehlungen

Medienkompetenz fördern. Eine Förderung der Medienkompetenz könnte negativen Auswirken vorbeugen. Dazu gehört sowohl die Fähigkeit, die Echtheit von Texten sowie Audio-, Bild- und Videoaufnahmen einschätzen zu können (also etwa die Beurteilung der (Ursprungs-)Quelle oder die Suche nach ergänzenden Informationen), als auch ein verantwortungsvoller Umgang bei der Erzeugung und Verbreitung veränderter Aufnahmen.

Plausibilität nicht mit Realität verwechseln. Auch bei Anwendungen ohne Täuschungsabsicht, also beispielsweise der Rauschentfernung bei Bildern, handelt es sich letztlich um Manipulation. Selbst wenn die Ergebnisse überzeugend sind, gibt es keine Garantie, dass tatsächlich die Realität abgebildet wird. Dies ist zum Beispiel bei synthetischen Daten oder der medizinischen Bildgebung relevant.

Gemeinnützige Anwendungen fördern. Die neuen Möglichkeiten zur Generierung und Veränderung von Text-, Audio-, Bild- und Videomaterial sollten nicht auf ihr Gefahrenpotenzial reduziert werden. Diese Möglichkeiten könnten zu Fortschritten in der Medizin, der Astronomie und möglicherweise auch weiteren gemeinnützigen Bereichen führen.

Aktuelle Methoden zur Feststellung der Authentizität nutzen. Ähnlich wie bei IT-Sicherheit sollte auf die Aktualität der Methoden zur Enttarnung von Deepfakes geachtet werden. Zwar wird auf neue Methoden wiederum mit verbesserten Fälschungen reagiert, aktuelle Methoden könnten aber zumindest Deepfakes, die nicht mehr dem Stand der Technik entsprechen, enttarnen.

Beispiel für die Anwendung von Neural Style Transfer.

Geografische Verortung

In den USA wurde schon früh zum Innovationsfeld »Künstlicher Realismus« geforscht und beim Indikator für die Qualität der Forschung ist ein deutlicher Vorsprung erkennbar. Doch bei der Quantität belegt mittlerweile die VR China den Spitzenplatz. Besonders die VR China, die USA und das Vereinigte Königreich sind stark vernetzt. Deutschland ist nach dem Vereinigten Königreich das stärkste europäische Land.

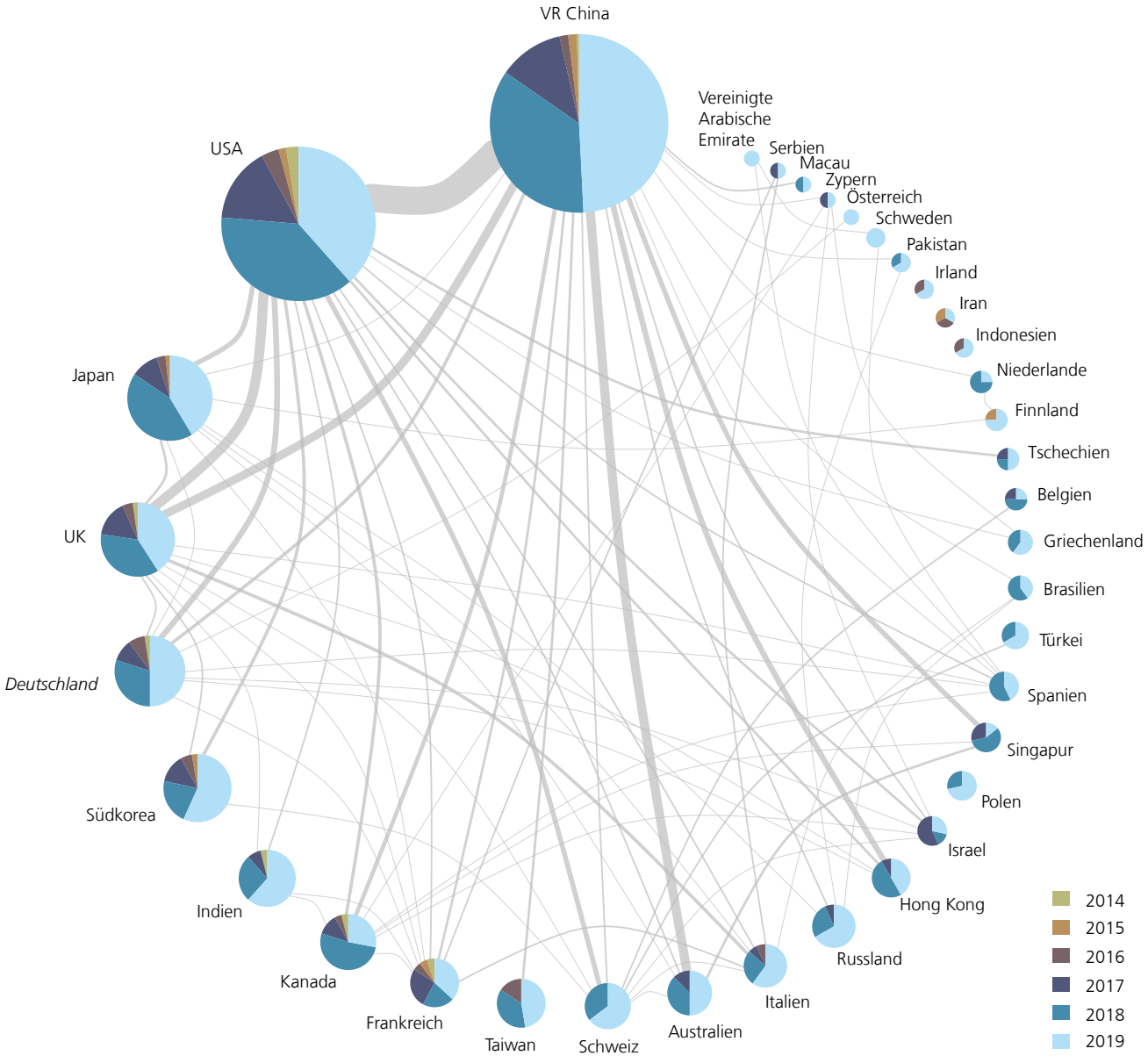


Abb. 16: Geografische Verteilung der Publikationshäufigkeit für das Innovationsfeld künstlicher Realismus. Publikationsanzahl: 2 bis 254. Gemeinsame Publikationsanzahl: 1 bis 34. Erhebung vom 10.11.2019.

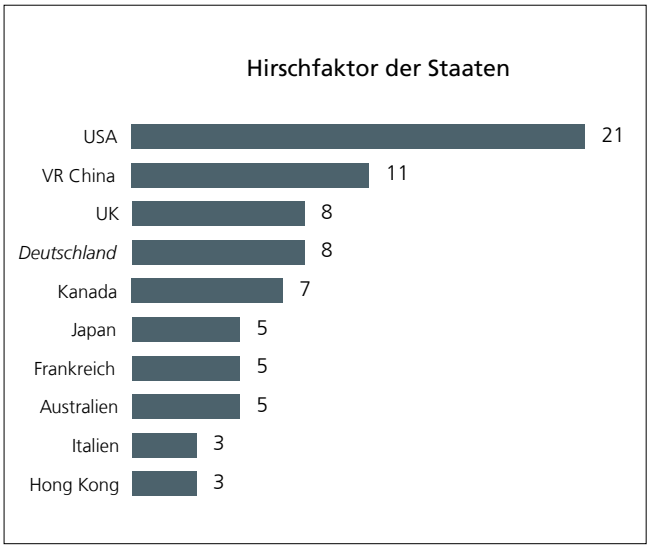


Abb. 17: Geografische Verteilung des Hirschfaktors für das Innovationsfeld künstlicher Realismus. Erhebung vom 10.11.2019.

Wahrnehmung in Wissenschaft und sozialen Medien

Durch die Hashtags in den sozialen Medien werden Medien, Fälschung und gezielte Desinformation betont. Die wissenschaftliche Diskussion wird hingegen vor allem durch die verschiedenen technischen Ansätze geprägt. Die Verwendung künstlicher Intelligenz kommt sowohl in den sozialen Medien als auch der Wissenschaft häufig zur Sprache.

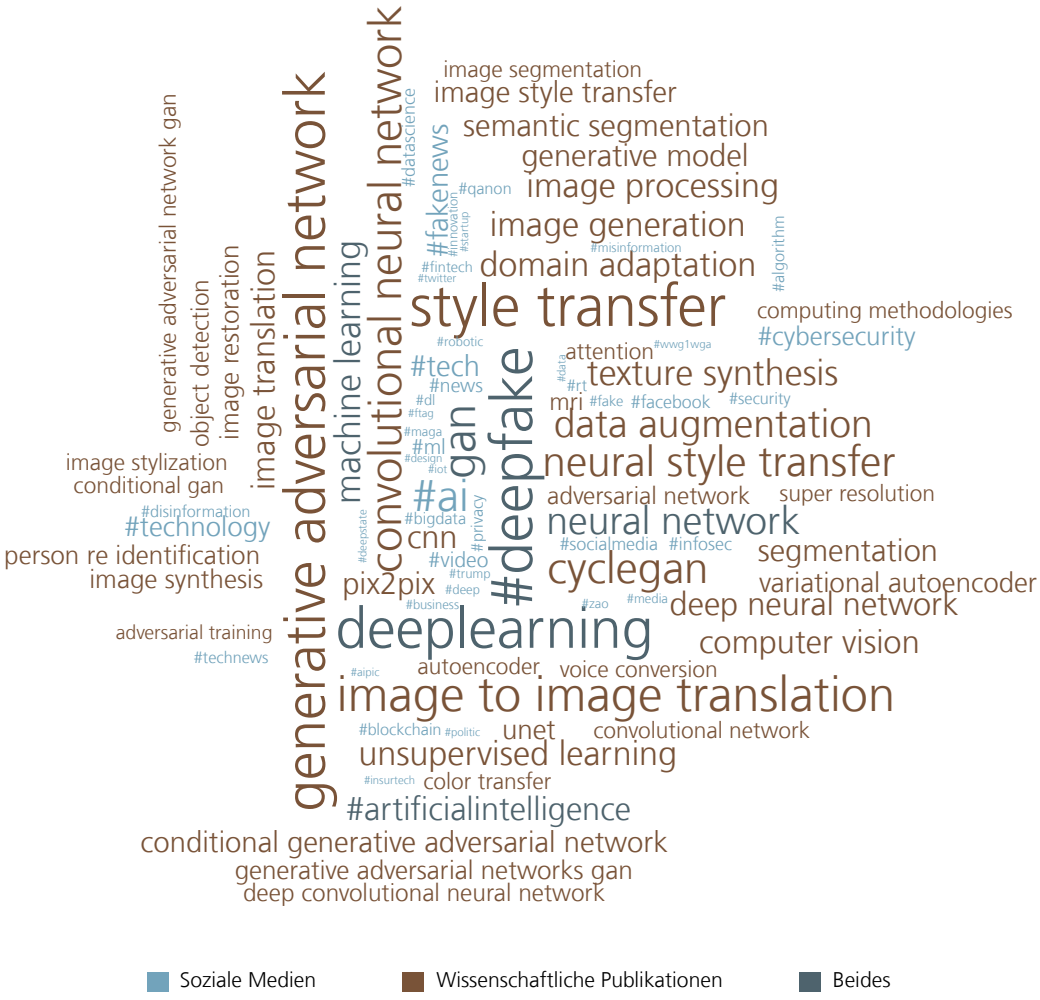
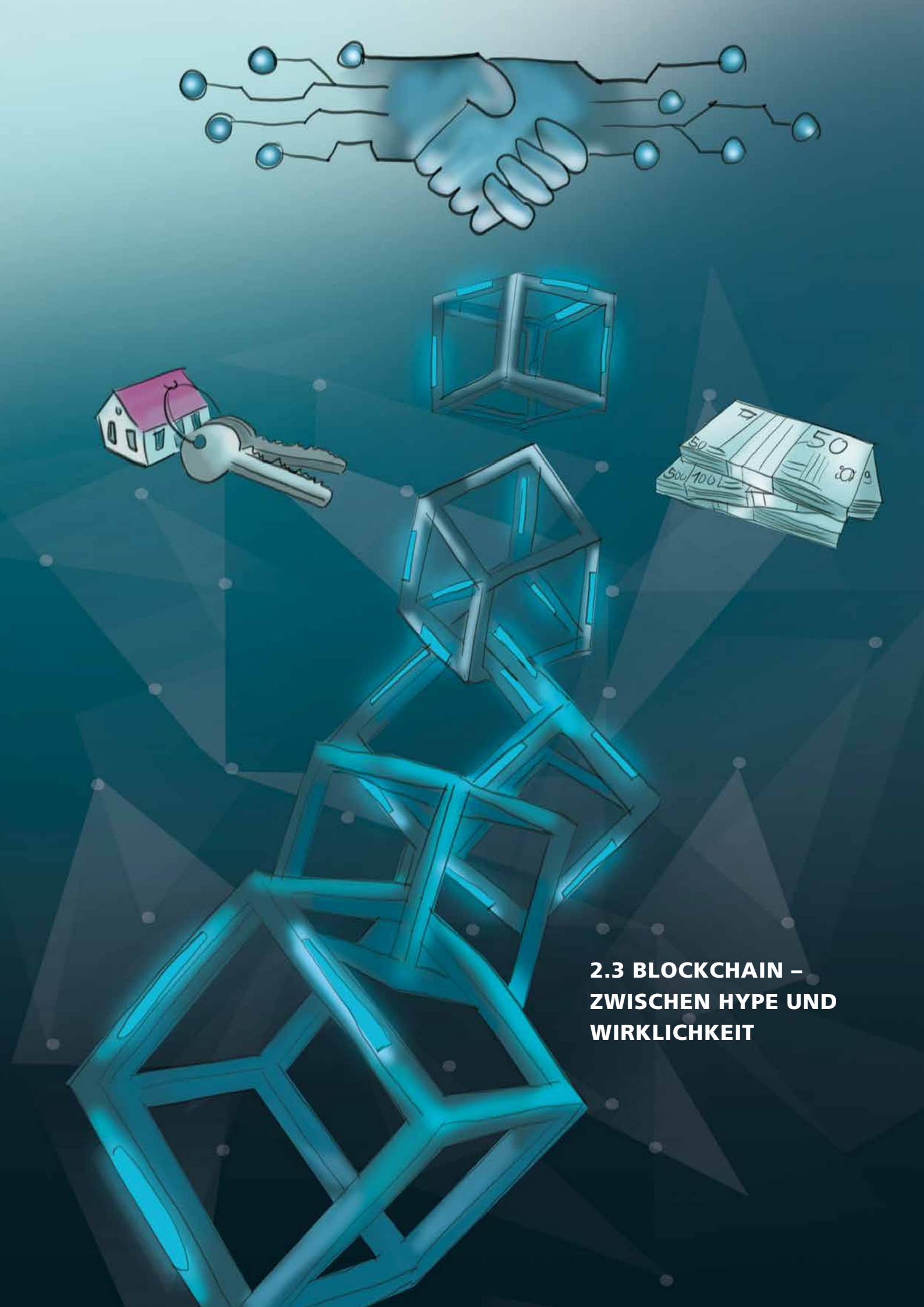


Abb. 18: Wordcloud für das Innovationsfeld künstlicher Realismus. Die häufigsten Hashtags und Schlüsselwörter für die Jahre 2018 und 2019. Erhebung vom 15.11.2019.



2.3 BLOCKCHAIN – ZWISCHEN HYPE UND WIRKLICHKEIT

Als 2008 die Kryptowährung Bitcoin entstand[35], war das ambitionierte Ziel, eine digitale Währung zu schaffen, die gänzlich ohne zentrale Instanzen wie beispielsweise Banken auskommt. Damit in einem solchen System Transaktionen trotzdem manipulationssicher und korrekt ablaufen, werden sie im Konsens zwischen allen Teilnehmenden überprüft, durchgeführt und nachvollziehbar protokolliert. Was aufwändig klingt, wird weitestgehend automatisiert. Durch eine geschickte Kombination aus Wettbewerb, Kryptografie und Transparenz bietet die Blockchain hierfür die technologische Basis.[36] Auch wenn durch Bitcoin zumindest bislang keine Bank abgelöst wurde, so hat die zugrunde liegende Technologie doch einen globalen Hype ausgelöst.

Während der Hype mittlerweile wieder etwas nachlässt, genießt die Blockchain bis in die Politik hinein noch immer hohe Aufmerksamkeit. Im Herbst 2019 hat die Bundesregierung eine eigene Blockchain-Strategie vorgelegt.[37] Darin werden fünf Handlungsfelder identifiziert und 44 Maßnahmen beschrieben, mit denen die Technologie im Wesentlichen erprobt werden soll.

Ein Blick in die vielfältigen Anwendungsgebiete zeigt die unterschiedlichen Perspektiven auf die Technologie. Mal wird sie als Kryptowährung gesehen, mal als Register (Datenbank) genutzt und mal als verteilter Computer verstanden, auf dem kleine Programme (sogenannte Smart Contracts) ablaufen.

Allgemein wird zwischen öffentlichen und privaten Blockchain-Lösungen unterschieden. Während erstere für jede:n frei einsehbar und zugänglich sind, werden private Blockchain-Lösungen vor allem im Verwaltungs- oder Unternehmensbereich eingesetzt. Private Blockchain-Lösungen werden oftmals auch als Konsortial-Blockchain bezeichnet. Zwischen öffentlichen und privaten Blockchain-Lösungen finden sich zudem weitere Abstufungen, bspw. in Bezug auf die Schreib- und Leserechte für Transaktionen. In der Regel werden hierbei »permissioned« und »permissionless« Blockchains unterschieden. Während bei »permissioned« Blockchains nur berechtigte Teilnehmer eines Netzwerkes Transaktionen veranlassen dürfen, ist dies bei »permissionless« Blockchains allen Teilnehmern gestattet. Diese Einordnung hilft insbesondere bei der Auswahl geeigneter Bausteine für eine Blockchain-Lösung. Klassische Kryptowährungen sind beispielsweise in der Regel als öffentliche und »permissionless« Blockchains konzipiert. In diesem Fall sind die kryptografischen Verfahren, der Konsensmechanismus sowie die technischen Lösungen zur Dezentralität von entscheidender Bedeutung. Bei privaten Blockchains wiederum können andere Aspekte im Vordergrund stehen, wie beispielsweise die Authentizität der Teilnehmer.

[35] Nakamoto, S. (2008).

[36] Welzel, C.; Eckert, K.; Kirstein, F.; Jacumeit, V. (2017).

[37] Siehe: https://www.blockchain-strategie.de.

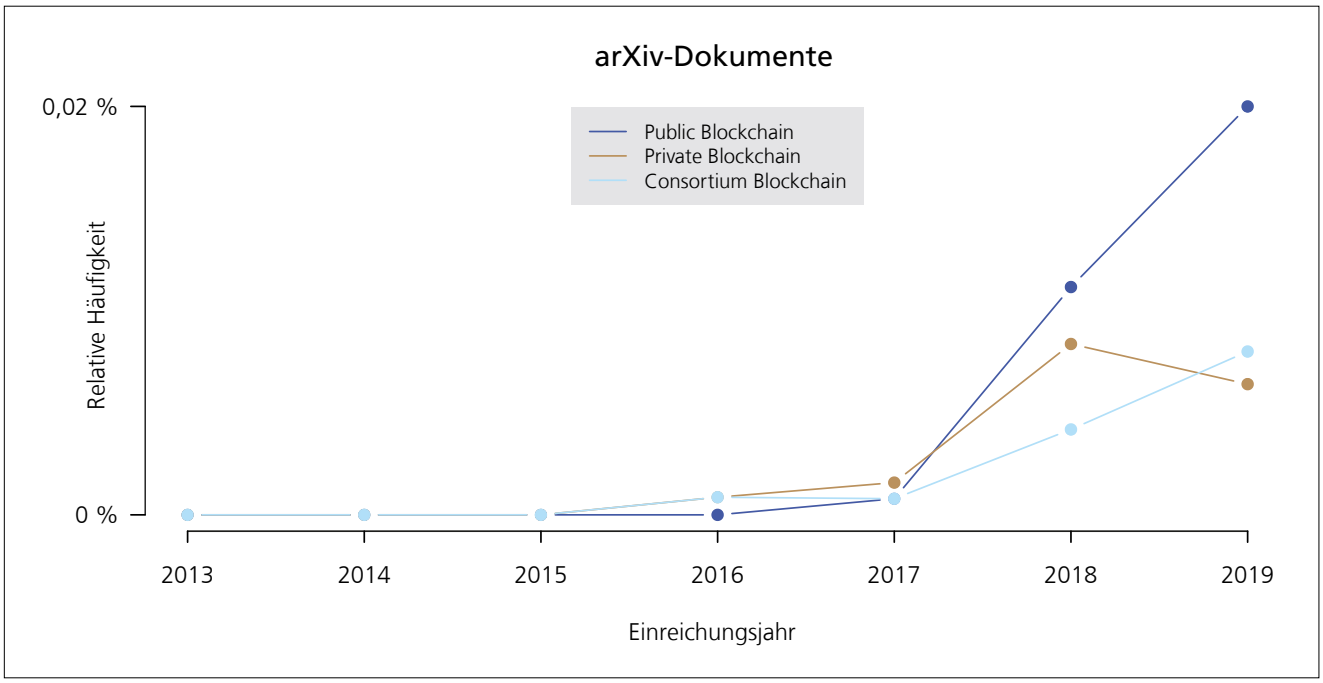


Abb. 19: Trendverlauf für verschiedene Blockchaintypen nach ihrem Vorkommen im Titel oder Abstract. Erhebung vom 18.10.2019.

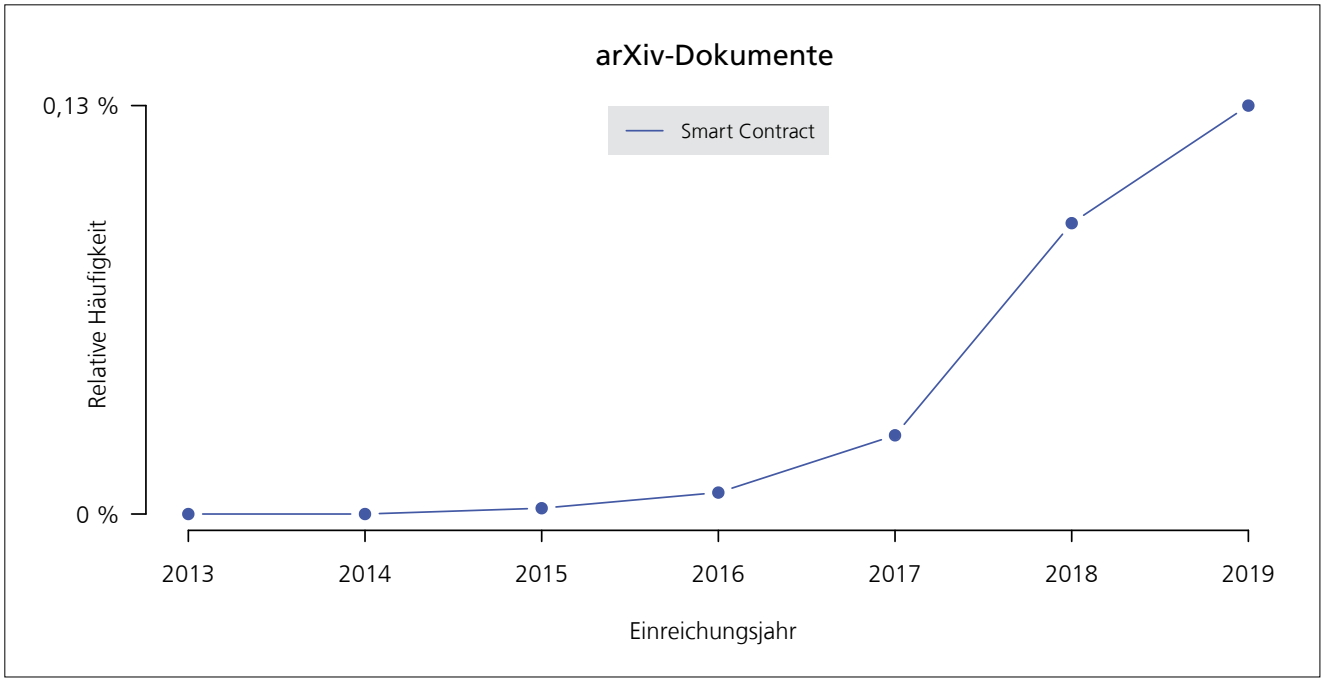


Abb. 20: Trendverlauf für Smart Contracts nach Vorkommen im Titel oder Abstract. Erhebung vom 18.10.2019.

Smart Contracts

Oftmals sind Transaktionen an bestimmte Vorbedingungen geknüpft, lösen weitere Aktivitäten aus oder sind in komplexe Abläufe eingebunden. Für diese Fälle wurden sogenannte Smart Contracts entwickelt.[38] Sie ermöglichen es, komplexe Transaktionen und deren Validierungsprozesse präzise zu beschreiben und zu programmieren. In diesem Fall fungiert ein Blockchain-Netzwerk als verteilter Computer, auf dem die Programme und Validierungsregeln ausgeführt werden. Wesentlicher Treiber hinter dieser Entwicklung ist das Blockchain-Netzwerk Ethereum.[39] Der Name Smart Contract vermittelt bereits, dass das komplette Regelwerk eines Vertrags abgebildet wird, sodass – zumindest in der Theorie – keine weitere Interaktion zwischen den Beteiligten notwendig ist.[40]

Smart Contracts werden als eigenständige, zustandsbehaftete, adressierbare Objekte betrachtet. Eine Zustandsänderung während der Ausführung eines Smart Contracts wird über Ereignisse ausgelöst. So könnte bspw. eine Versicherungssumme automatisch ausgezahlt werden, wenn ein definierter Schadensfall eintritt. Die Ausführung eines Smart Contracts erfolgt in einzelnen Schritten. Die dafür notwendige Rechenleistung wird über die Blockchain-eigene Kryptowährung abgerechnet. Jeder Smart Contract muss daher mit einem entsprechenden Budget ausgestattet werden. Durch die Festlegung einer maximalen Anzahl von Ausführungsschritten je Smart Contract kann verhindert werden, dass zu viel Budget verbraucht wird. Smart Contracts bieten einen vielversprechenden Ansatz, um bestimmte, insbesondere überschaubare und automatisierbare Prozesse vollautomatisiert ablaufen zu lassen.

Alternative Konsensverfahren

Ein häufig thematisiertes Problem aktueller Blockchain-Implementationen ist ihr enormer Energieverbrauch.[41] Dieser ist vorrangig durch das aufwändige Konsensverfahren namens Proof-of-Work bedingt.[42] Dabei versuchen die Betreiber der Knoten (Teilnehmer) des Blockchain-Netzwerks im Wettbewerb ein kryptografisches Rätsel zu lösen. Wer dieses zuerst löst, kann ausstehende Transaktionen in Form eines neuen Blocks zur Blockchain hinzufügen, wofür es im Gegenzug als Belohnung eine definierte Menge der jeweiligen Kryptowährung gibt. Da die Knoten mit höherer Rechenleistung eine größere Chance haben, das Rätsel zuerst zu lösen, besteht ein starker Anreiz, in mehr Rechenleistung zu investieren. Der Schwierigkeitsgrad des kryptografischen Rätsels steigt mit der verfügbaren Rechenleistung. Diese Spirale führt zu einem steigenden Energieverbrauch für den Betrieb der Kryptowährung.

[38] Zakhary, V.; Agrawal, D.; El Abbadi, A. (2019).

[39] Siehe: https://www.ethereum.org/.

[40] Green, S. (2019).

[41] Das, D.; Dutta, A. (2019).

[42] Li, J.; Li, N.; Peng, N.; Cui, H.; Wu, Z. (2019).

Daher wird aktuell intensiv an alternativen Konsensmechanismen geforscht. Am weitesten fortgeschritten ist derzeit das Verfahren Proof-of-Stake, welches bereits in einigen Blockchain-Implementierungen Anwendung findet.[43] Hierbei ist nicht mehr die Rechenleistung für die Erzeugung neuer Blöcke ausschlaggebend, sondern der Anteil am Vermögen der jeweiligen Kryptowährung. Dem zugrunde liegt die Vermutung, dass Eigentümer:innen mit vielen Anteilen einer Kryptowährung ein geringeres Manipulationsinteresse haben. Ob dieser Ansatz jedoch für alle Anwendungsfälle geeignet ist, sei dahingestellt, denn er zementiert existierende Machtverhältnisse und verstärkt das Ungleichgewicht zwischen Teilnehmer:innen mit geringen und jenen mit hohen Anteilen.

Eine Weiterentwicklung versucht, dieses Ungleichgewicht etwas aufzulösen. Beim Delegated Proof-of-Stake-Verfahren[44] können mehrere Teilnehmer:innen ihre Anteile zusammenlegen, um so ihre Chancen zu vergrößern.

Eine Weiterentwicklung von Proof-of-Stake stellt auch das Konsensverfahren Proof-of-Authority dar. Anstelle der Anteilsmenge an einer Kryptowährung spielt hier die Reputation der Teilnehmer eine wesentliche Rolle. Diese wird in einem definierten Prozess durch einen Validierer gegeben. Welche Faktoren für die Reputation ausschlaggebend sind, ist dabei von essenzieller Bedeutung. Ein anderer Ansatz wiederum basiert darauf, dass jedem Knoten (Teilnehmer) per Zufallsprinzip eine Wartezeit zugeordnet wird, nach der dieser einen neuen Block hinzufügen kann und dafür entsprechend entlohnt wird. Dieses Verfahren wird auch als Proof-of-Elapsed-Time bezeichnet. Sowohl Proof-of-Authority als auch Proof-of-Elapsed-Time benötigen zentrale Instanzen oder Intermediäre, beispielsweise einen Zeitgeber oder einen Validierer. Sie eignen sich daher eher für private Blockchain-Anwendungen, bei denen – im Gegensatz zu öffentlichen Blockchain-Anwendungen – die vollständige Dezentralisierung oftmals nicht das maßgebende Prinzip darstellt.[45]

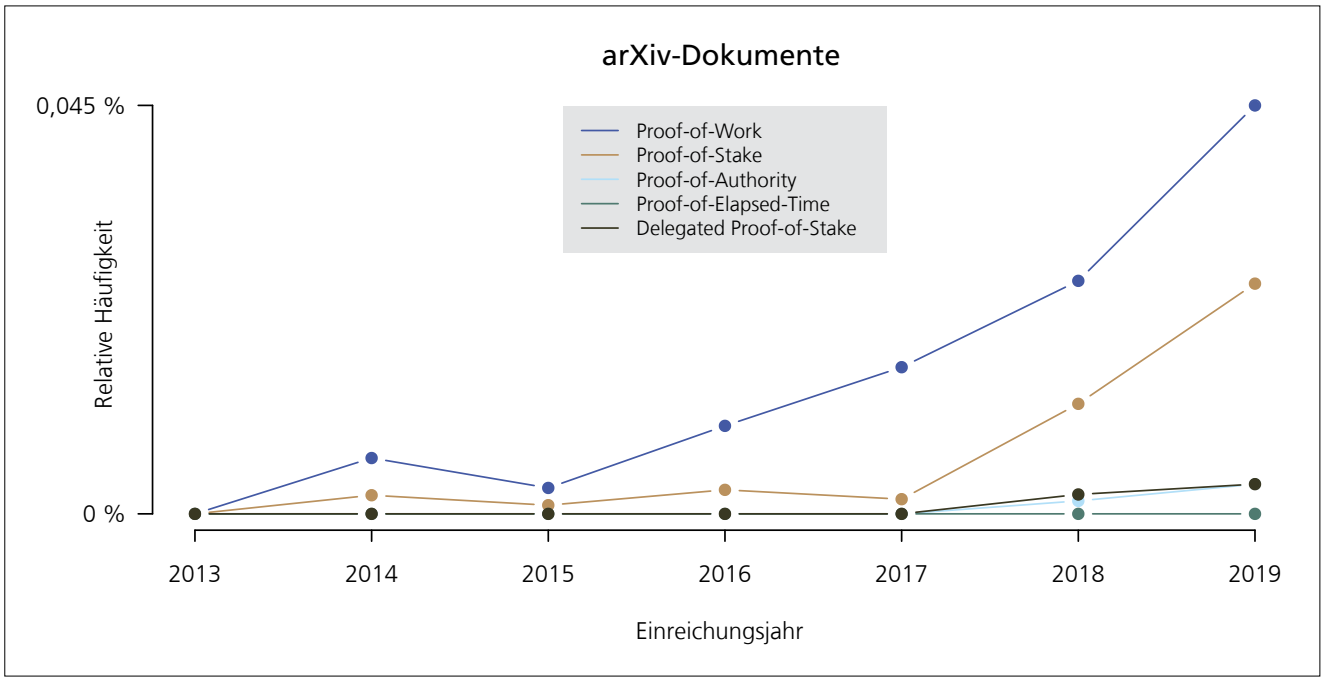


Abb. 21: Trendverlauf für verschiedene Konsensverfahren nach ihrem Vorkommen im Titel oder Abstract. Erhebung vom 18.10.2019.

[43] Saleh, F. (2019).

[44] Nguyen, C. T.; Hoang, D. T.; Nguyen, D. N.; Niyato, D.; Nguyen, H. T.; Dutkiewicz, E. (2019).

[45] Bano, S.; Sonnino, A.; Al-Bassam, M.; Azouvi, S.; McCorry, P.; Meiklejohn, S.; Danezis, G. (2017).

Anwendungsfelder und Beispiele

Eine Blockchain kann überall da nützlich sein, wo viele Akteure über Organisationsgrenzen hinweg koordiniert werden müssen, bzw. sich die Akteure nicht direkt kennen oder Transaktionen sicher zu dokumentieren sind. Wenn zusätzlich Manipulationsanreize bestehen oder das Vertrauen zwischen den Akteuren fehlt, bietet die Blockchain einen technologischen Lösungsansatz.

Mittlerweile wird die Blockchain-Technologie daher weit über den Finanzsektor hinaus diskutiert, erprobt oder bereits eingesetzt. Anforderungen wie Integrität, Transparenz und Dezentralität spielen in vielen Bereichen eine wichtige Rolle. So kommt die Blockchain etwa bei der Absicherung von Produktions- und Logistikketten zum Einsatz.[46] In einer globalisierten Welt mit komplexen Zuliefer- und Fertigungsnetzwerken kann so jederzeit nachvollzogen werden, in welchem Produktionsstatus sich ein Produkt aktuell befindet oder wer wann an der Herstellung beteiligt war. Dies ist beispielsweise bei der Lebensmittelproduktion wichtig.

Im Bildungsbereich kann die Integrität elektronischer Zeugnisse oder anderweitiger Ausbildungsnachweise über eine Blockchain abgesichert werden.[47] Auch in der Verwaltung wird der Einsatz der Blockchain-Technologie diskutiert und teilweise erprobt. Ein vielbeachtetes Beispiel ist hier etwa das Bundesamt für Migration und Flüchtlinge, welches derzeit den Einsatz der Blockchain-Technologie im Asylprozess pilotiert.[48] Auch auf Ebene der Bundesländer gibt es erste Pilotprojekte. In Nordrhein-Westfalen wird mit dem Projekt GovChain NRW[49] eine landesweite Blockchain-Infrastruktur über die kommunalen Rechenzentren aufgebaut. In Thüringen wiederum wird aktuell das Konzept der LifeChain in einer Testumgebung erprobt. Dabei können Bürger:innen Verwaltungsangelegenheiten mittels eines Blockchain-basierten Bürgerkontos nutzen.[50] Darüber hinaus sind Szenarien zur Integritätssicherung amtlicher Dokumente oder dem Aufbau dezentraler Register denkbar.[51]

Weitere Anwendungen finden sich im Energiesektor. Hier entstehen basierend auf Blockchain Handelsplattformen für den direkten Austausch zwischen Stromerzeuger:innen und Verbraucher:innen.[52,53] Darüber hinaus kann die Blockchain auch für die Steuerung der Energieinfrastruktur eingesetzt werden, bspw. für den Flexibilitätenhandel.[54]

Auswirkungen, Chancen und Risiken

Viele der Anwendungsfälle können auch mit alternativen Technologien umgesetzt werden. Die Blockchain tritt daher meist in Konkurrenz zu etablierten technischen Lösungen auf, wie verteilten Datenbanksystemen oder elektronischen Signaturen. Die Gründe für die Wahl der Blockchain-Technologie sind sehr unterschiedlich. Mal ist es die Dezentralität der Lösung, mal ist eine Blockchain-Lösung kostengünstiger, mal überwiegen Marketing- und Kommunikationsaspekte.

Trotz der vielen diskutierten Anwendungsfälle stellt Bitcoin bis heute noch immer das größte Blockchain-System dar. Auch hierfür gibt es viele Gründe. Die Technologie ist in vielen Bereichen noch nicht ausgereift und oftmals sehr komplex, nicht zuletzt deswegen, weil sich an einer produktiven Umsetzung viele Parteien beteiligen müssen. Hinzu kommen rechtliche Fragen. Die Blockchain-Technologie ist darauf ausgelegt, dass allen Teilnehmenden jederzeit alle Informationen zur Verfügung stehen und Transaktionsinformationen im Nachhinein nicht geändert oder gelöscht werden können. Dies steht im direkten Gegensatz zu den Artikeln 16 und 17 der europäischen Datenschutzgrundverordnung.[55,56]

Aktuell findet im Blockchain-Umfeld eine intensive Forschung und Entwicklung statt. Daher ist es schwer vorherzusagen, wie sich die Technologie weiterentwickeln wird. Als Infrastruktur-Technologie wird sie in Zukunft vermutlich eher unsichtbarer werden. Ihr Charakteristikum, die dezentrale Koordination unterschiedlicher Akteure, ist jedoch für viele Anwendungsbereiche interessant.

[46] Chang, Y.; Iakovou, E.; Shi, W. (2019).

[47] Bessa, E. E.; Martins, J. S. B. (2019).

[48] Siehe: http://www.bamf.de/DE/DasBAMF/BAMFdigital/Blockchain/blockchain-node.html.

[49] Siehe: https://govchain-blog.de/govchain-nrw-start-des-reallabors-fuer-eine-government-blockchain-infrastructure-in-nrw/.

[50] Siehe: https://www.bundesdruckerei.de/de/Newsroom/Pressemitteilungen/Sichere-Loesung-fuer-Buergerkonten-nach-dem-Once-Only-Prinzip.

[51] Initiative Blockchain in der Verwaltung (2019).

[52] Hassan, N. U.; Yuen, C.; Niyato, D. (2019).

[53] Beispielsweise: https://www.lition.de/.

[54] Rangelov, D.; Tcholtchev, N.; Lämmel, P.; Schieferdecker, I. K. (2019).

[55] Siehe Europäische Datenschutz-Grundverordnung: https://eur-lex.europa.eu/legal-content/DE/TXT/?uri=CELEX:02016R0679-20160504.

[56] Finck, M. (2019).

Handlungsempfehlungen

Einsatz der Blockchain-Technologie kritisch prüfen. Vor dem Einsatz der Blockchain-Technologie sollte kritisch geprüft werden, inwieweit sie tatsächlich sinnvoll ist. Vielfach können bestehende Fragestellungen auch mit alternativen Technologien umgesetzt werden. In diesen Fällen hilft eine vergleichende Effizienzbetrachtung.

Governance-Regeln bereits bei der Konzeption definieren. Insbesondere öffentliche Blockchains zeichnen sich dadurch aus, dass sie nicht von einer zentralen Instanz gesteuert werden. Daher ist bei der Konzeption und Entwicklung ein besonderes Augenmerk auf die Governance zu legen, denn spätere Änderungen sind nur über sehr aufwändige Abstimmungen umsetzbar.

Klein anfangen und Erfahrungen sammeln. Um die Eignung und Relevanz der Blockchain-Technologie einzuschätzen, ist es notwendig, praktische Erfahrungen zu sammeln. Vom Proof of Concept über Prototypen bis hin zu Pilotprojekten gibt es viele Möglichkeiten, auch mit geringem Aufwand mögliche Anwendungsfälle zu erproben und so Erfahrungen aufzubauen und Best-Practice-Beispiele zu entwickeln.

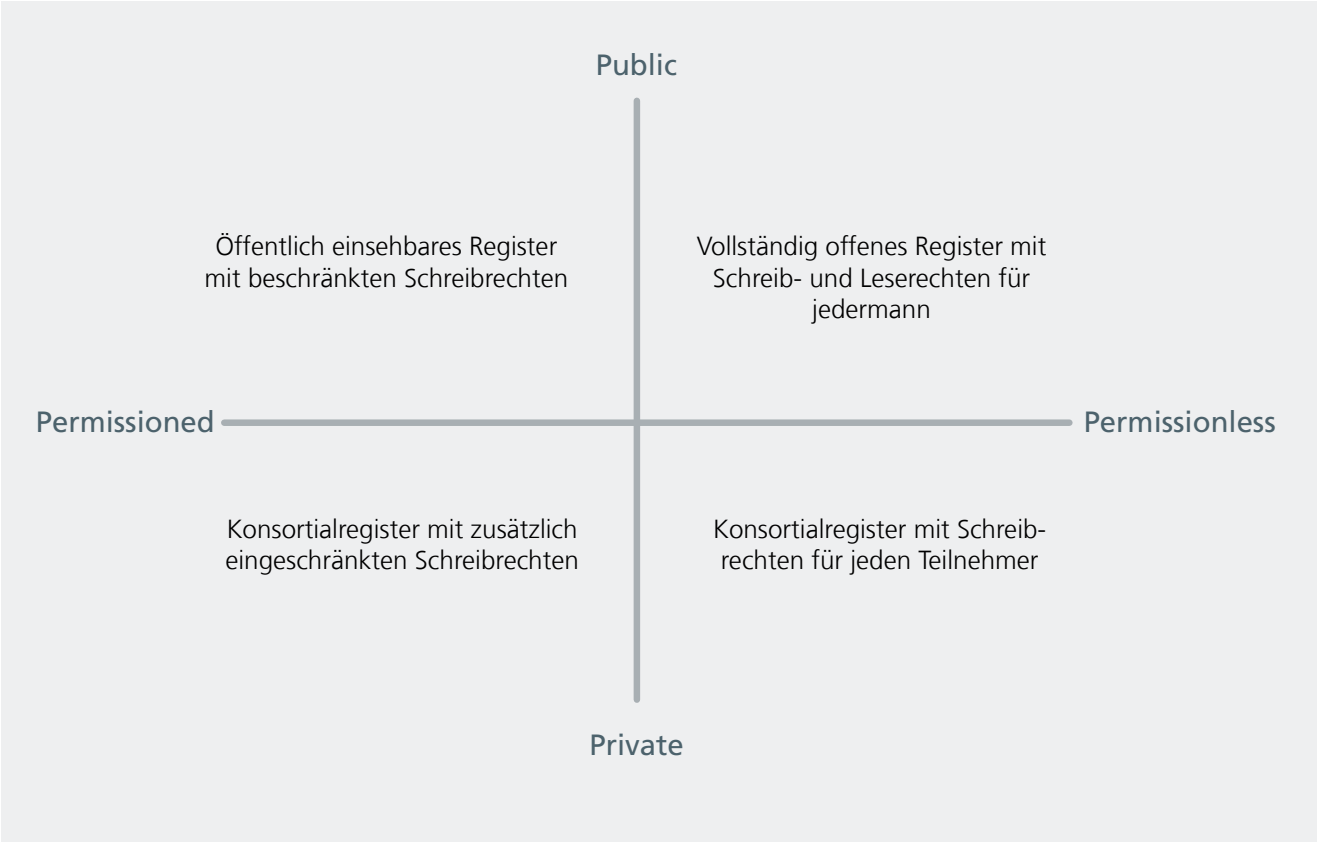


Abb. 22: Mögliche Blockchain-Lösungen

Geographische Verortung

Die USA und China führen die Forschung zum Innovationsfeld Blockchain an, Deutschland belegt beim Indikator für Quantität den fünften und beim Indikator für Qualität den sechsten Platz. Im Vergleich zu den anderen Innovationsfeldern fällt die starke internationale Vernetzung auf. Die deutliche Mehrheit der Publikationen stammt aus den Jahren 2018 und 2019.

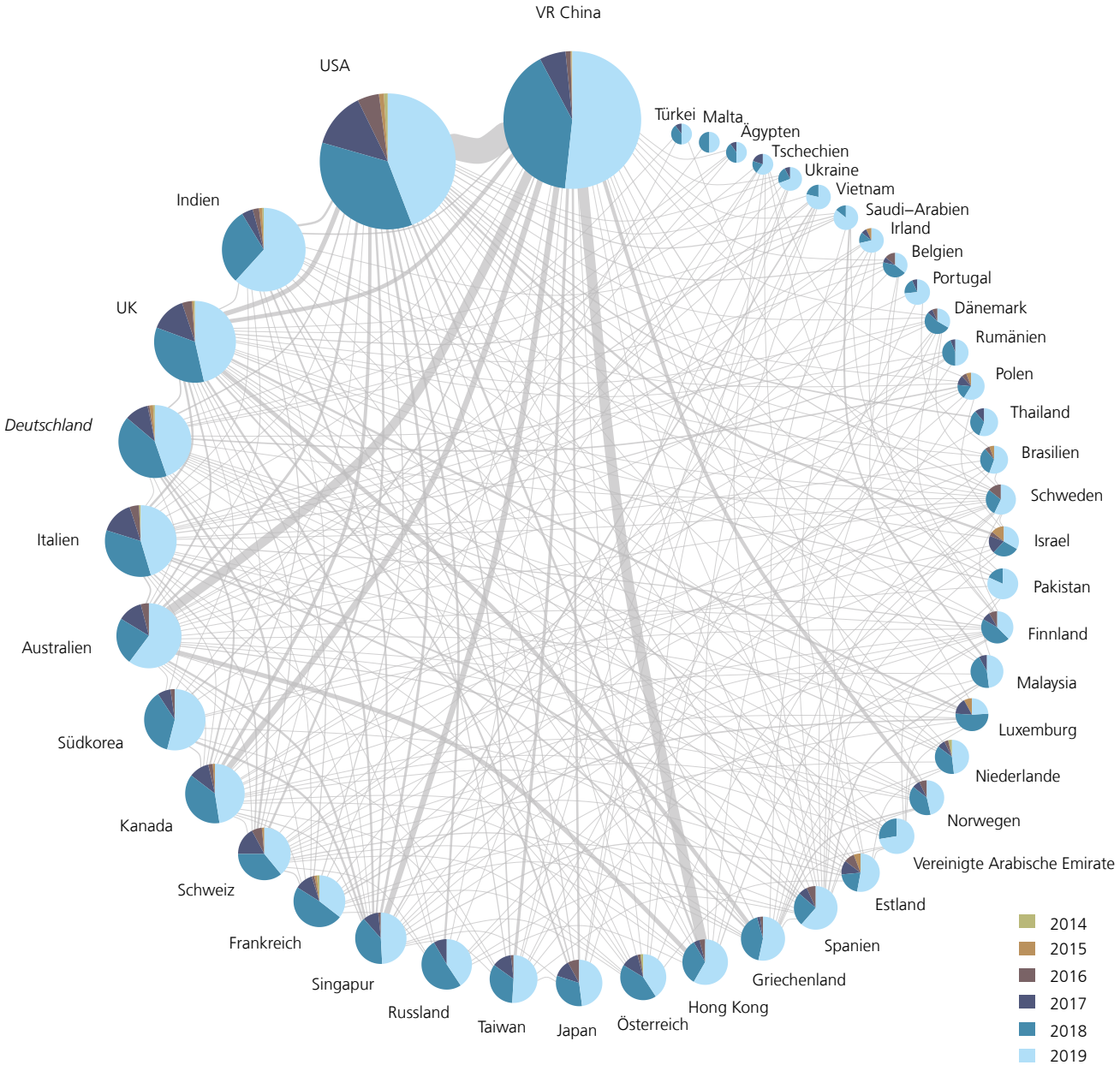


Abb. 23: Geografische Verteilung der Publikationshäufigkeit für das Innovationsfeld Blockchain. Publikationsanzahl: 10 bis 439. Gemeinsame Publikationsanzahl: 1 bis 69. Erhebung vom 10.11.2019.

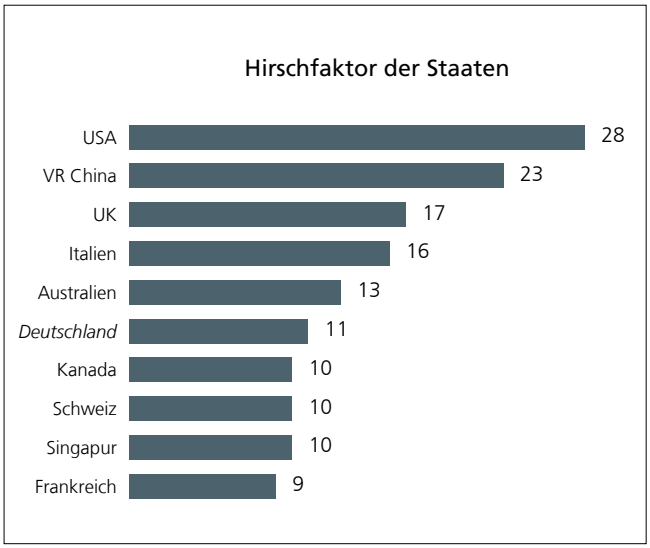


Abb. 24: Geografische Verteilung des Hirschfaktors für das Innovationsfeld Blockchain. Erhebung vom 10.11.2019.

Wahrnehmung in Wissenschaft und sozialen Medien

Während in den sozialen Medien der Fokus deutlich auf dem Finanzsektor liegt, indem primär über Kryptowährungen und Fintechs diskutiert wird, zeigt sich in der Analyse der wissenschaftlichen Publikationen ein diverseres Bild. In den Veröffentlichungen befassen sich die Autoren mit den Themen verteilte Register und alternative Konsensverfahren sowie weiteren Anwendungsgebieten, wie dem Internet der Dinge oder Logistikketten aber auch mit übergreifenden Themen wie Datenschutz und Vertrauen.

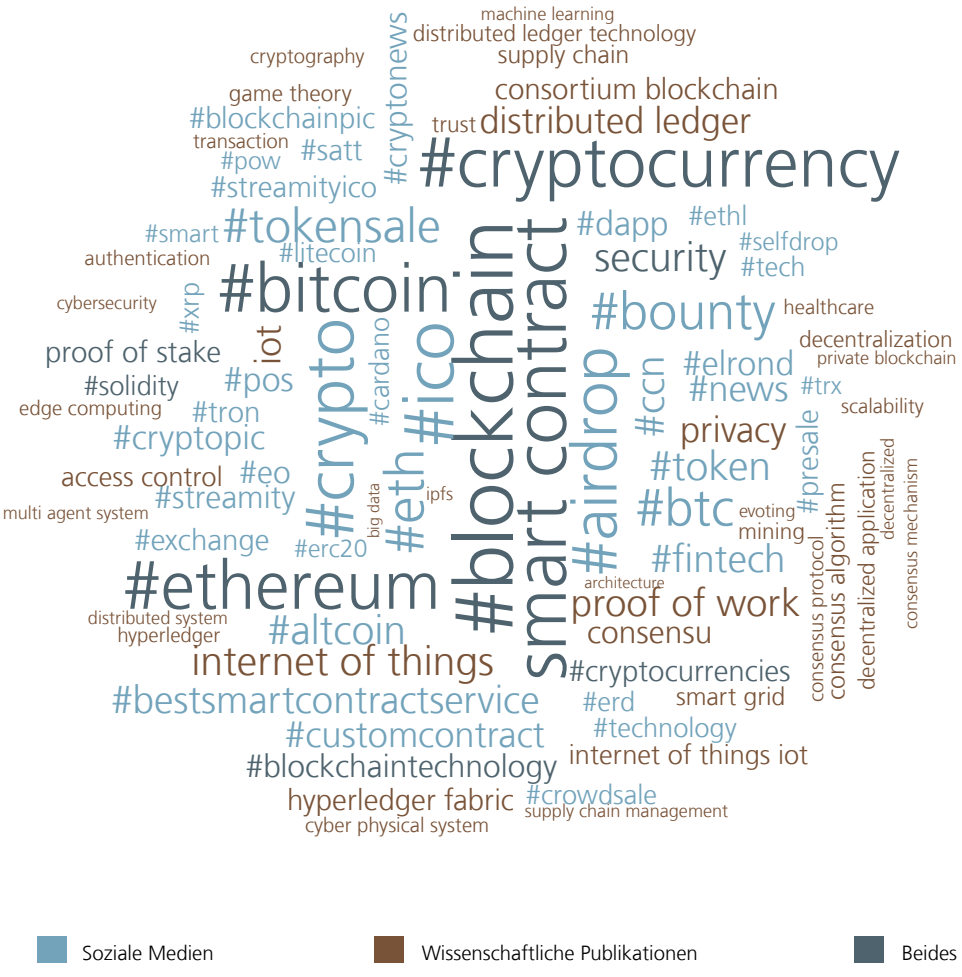


Abb. 25: Wordcloud für das Innovationsfeld Blockchain. Die häufigsten Hashtags und Schlüsselwörter für die Jahre 2018 und 2019. Erhebung vom 17.11.2019.

2.4 DIE ACHILLESFERSE DER KI?

Wenn heutzutage von künstlicher Intelligenz gesprochen wird, sind zumeist Systeme gemeint, die anhand von auf Machine Learning basierenden Modellen Entscheidungen treffen. Dabei kommen oftmals künstliche neuronale Netze zum Einsatz. Anwendungsfelder, in denen KI-Technologie bereits heute eingesetzt wird und weiter an Bedeutung gewinnt, sind zum Beispiel Sprachassistenten, autonome Fahrzeuge, Malware-Erkennung, Empfehlungsdienste, Contentfilter, die Auswertung medizinischer Bilder und biometrische Sicherheitssysteme. Aufgrund der zunehmenden Verbreitung und der teilweisen Kritikalität der Anwendungsfelder ist absehbar, dass KI-Systeme ein beliebtes Angriffsziel sein werden. Insbesondere im Hinblick auf den Einsatz von KI bei kritischen Systemen ist es daher besorgniserregend, dass in den letzten Jahren erhebliche Schwachstellen und variantenreiche Angriffsmöglichkeiten entdeckt wurden.

Angriffsarten

Angriffe auf KI Systeme lassen sich anhand verschiedener Kriterien unterscheiden, etwa anhand

- des Wissens der Angreifenden,
- des Ziels der Angreifenden oder
- des Stadiums des Machine-Learning-Modells (Training oder Anwendung).

Bei White-Box-Angriffen kennen die Angreifenden die innere Funktionsweise des Machine-Learning-Modells, bei Black-Box-Angriffen hingegen nicht. Weiter lassen sich gezielte und ungezielte Angriffe (Targeted Attack und Non-Targeted Attack) unterscheiden. Bei einem ungezielten Angriff sollen die Eingaben der Angreifenden inkorrekt verarbeitet werden. Bei einem gezielten Angriff soll darüber hinaus eine bestimmte Ausgabe erreicht werden. Vereinfachend[57] lässt sich die Entwicklung eines Machine-Learning-Systems in zwei aufeinanderfolgende Stadien unterteilen: Training und Anwendung. Für beide Phasen existieren Angriffs- und Verteidigungsmöglichkeiten, die derzeit rasant weiterentwickelt werden.

Data Poisoning

Bei Data Poisoning handelt es sich um die gezielte Beeinträchtigung der Integrität von Daten. Mittels Data Poisoning können Machine-Learning-Modelle während der Trainingsphase angegriffen werden. Dazu werden »vergiftete« Beispiele in den Trainingsdatenbestand eingeschleust. Bei solchen »vergifteten« Beispielen handelt es sich etwa um geschickt veränderte oder falsch bezeichnete Daten. Es gibt verschiedene Wege, Daten illegitim und unbemerkt zu modifizieren. Falls der Angriff durch einen Insider erfolgt oder die Daten (beispielsweise im Rahmen eines Open-Source-Projekts) auf öffentlich zugänglichen Servern gehostet werden, besteht sogar ein direkter Zugang.[58]

[57] Der Übergang zwischen Anwendung und Training ist mitunter fließend. Zur Anwendung gehört zudem auch die Testphase vor dem tatsächlichen Einsatz eines ML-Modells.

[58] Chen, X.; Liu, C.; Li, B.; Lu, K.; Song, D. (2017).

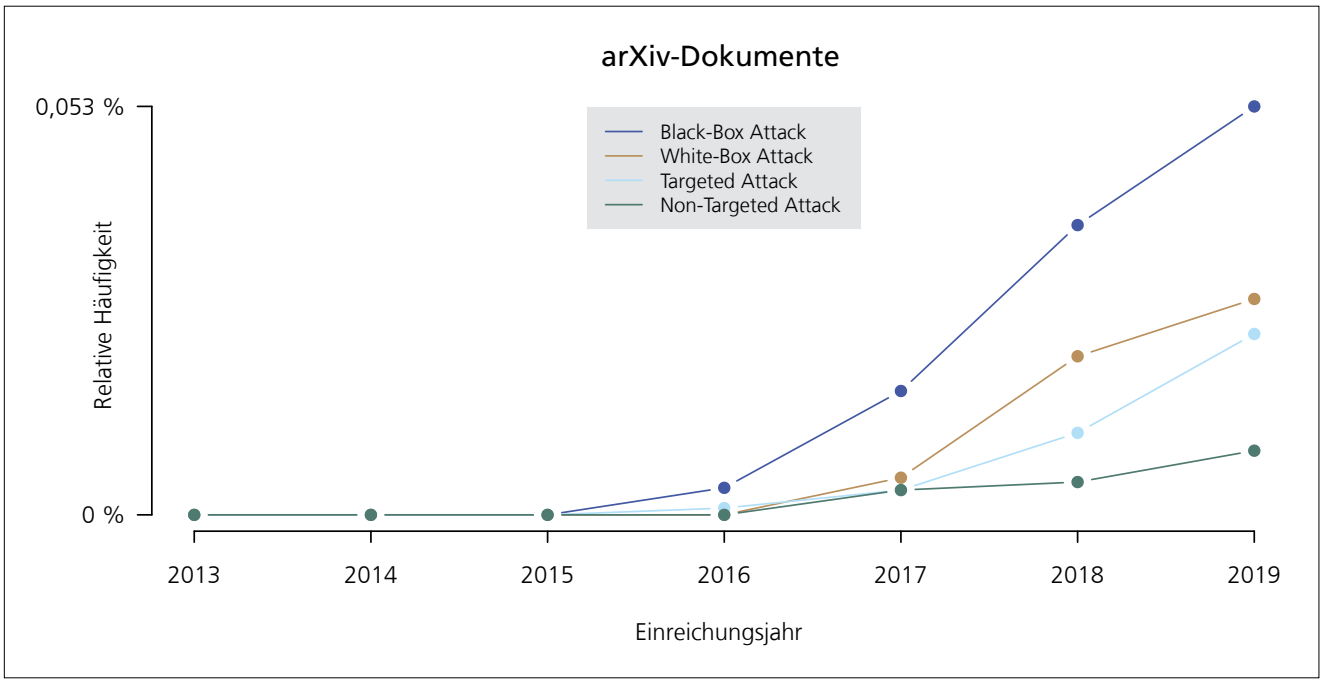


Abb. 26: Trendverlauf für verschiedene Angriffsarten nach ihrem Vorkommen im Titel oder Abstract. Erhebung vom 11.11.2019.

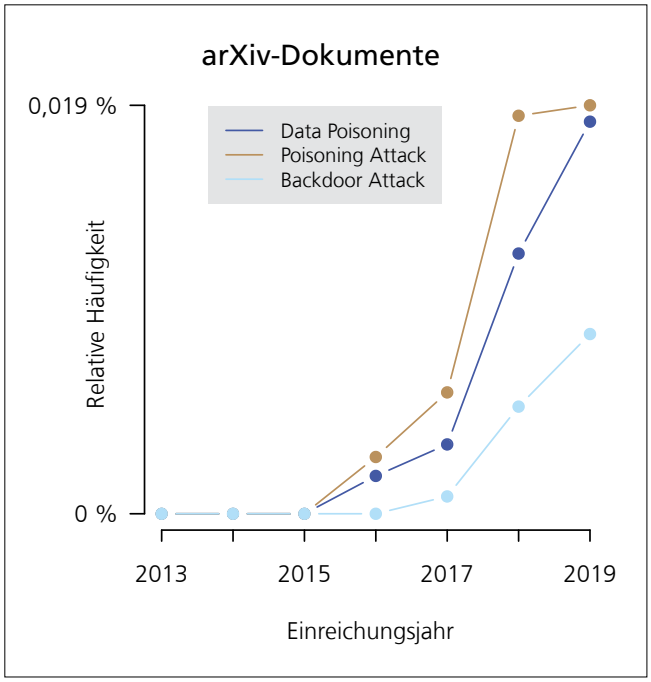


Abb. 27: Trendverlauf für Data Poisoning nach Vorkommen im Titel oder Abstract. Erhebung vom 11.11.2019.

Alternativ können sich Angreifer durch Ausnutzen von Netzwerkschwachstellen in der IT-Infrastruktur Zugang zu den Daten verschaffen. Wenn beispielsweise ein Modell einen Web Scraper[59] zur Sammlung von Daten einsetzt, kann es durch im Internet frei zur Verfügung gestellte »vergiftete« Daten angegriffen werden.[60] Der Trainingsaufwand für Machine-Learning-Modelle ist teilweise enorm und kann sich bei bestimmten Modellen sogar bei Ausführung auf modernster Parallelhardware im Bereich von Tagen und Wochen bewegen. Um in der Praxis den Trainingsaufwand zu reduzieren, wird daher gelegentlich auf vortrainierte Modelle zurückgegriffen, die entweder unmittelbar eingesetzt oder vorher einem anwendungsspezifischen Feintuning unterzogen werden, ggf. unter Nutzung zusätzlicher Trainingsdaten. Hier bietet sich für Angreifende die Möglichkeit, ein solches Modell unter anderem anhand einiger »vergifteter« Beispiele lernen zu lassen und es anschließend zur Verfügung zu stellen.[61] Bemerkenswert ist, dass schon sehr kleine Mengen an »vergifteten« Daten für erfolgreiche Poisoning Attacks ausreichen[62], was erfolgreiche Angriffe erleichtert.

Eine Poisoning Attack kann darauf abzielen, ein Modell insgesamt unbrauchbar oder weniger verlässlich zu machen, indem der Anteil der korrekten Entscheidungen reduziert wird. Bei einer Backdoor Poisoning Attack soll die allgemeine Funktionalität des Modells hingegen erhalten bleiben. Für gewisse Eingaben, etwa Daten, die mit einem bestimmten Muster versehen wurden, trifft das Modell jedoch zuverlässig die falschen Entscheidungen. Beispielhaft lässt sich dies an einem KI-System erklären, das Bilder von Hunden von Bildern von Katzen unterscheiden soll. Angenommen während der Trainingsphase werden korrekt bezeichnete Bilder von Hunden eingeschleust, die allesamt das gleiche unauffällige Muster am linken Bildrand enthalten. Unter allen Trainingsdaten befindet sich jedoch kein Katzenbild, das ebenfalls dieses Muster aufweist. Dadurch lernt die KI, dass das Muster sehr stark mit der Klassifikation »Hundebild« korreliert. Während der Anwendungsphase funktioniert die KI dann zuverlässig für zufällig ausgewählte Bilder von Hunden oder Katzen. Wird nun ein Katzenbild eingegeben, das mit dem Muster am linken Bildrand versehen worden ist, führt dies zu jedoch zur inkorrekten Klassifizierung als »Hundebild«, weil das Muster stärker als alle anderen Bildeigenschaften bewertet wird. Durch eine solche Hintertür, deren Vorhandensein im fertigen Modell überaus schwer festzustellen ist, besteht also die Möglichkeit, das Modell täuschen.

Eine erfolgreiche Poisoning Attack gelang z. B. beim Microsoft-Chatbot Tay. Nach weniger als 24 Stunden auf Twitter war es einigen Nutzer:innen gelungen, dem Bot rassistische Parolen beizubringen.[63]

Adversarial Examples

Adversarial Examples zielen darauf ab, ein bereits trainiertes Machine-Learning-Modell zu täuschen. Es handelt sich dabei um geringfügig veränderte Eingaben, die das Modell in ihrer ursprünglichen Form mit hoher Zuverlässigkeit korrekt verarbeitet. Die für den Menschen oft gar nicht wahrnehmbaren Änderungen, auch bekannt als Adversarial Perturbations, führen nun jedoch zu einer anderen und damit Fehlverarbeitung durch das Modell.

In der Fachliteratur werden Adversarial Examples zumeist in der Bilddatenklassifizierung diskutiert. Die Erstellung eines Adversarial Image ist in manchen Fällen schon durch Änderung des Werts eines einzigen von 227-mal-227 Bildpunkten möglich.[64] Doch auch für (durch Kameras erfasste) physische Objekte[65],

[59] Web Scraper sind Werkzeuge zur automatisierten Datensammlung bei Websites.

[60] Zhu, C.; Huang, W. R.; Shafahi, A.; Li, H.; Taylor, G.; Studer, C.; Goldstein, T. (2019).

[61] Gu, T.; Liu, K.; Dolan-Gavitt, B.; Garg, S. (2019).

[62] Steinhardt, J.; Koh, P. W.; Liang, P. (2017).

[63] Postinett, A. (2016).

[64] Su, J.; Vargas, D. V.; Kouichi, S. (2019).

[65] Athalye, A.; Engstrom, L.; Ilyas, A.; Kwok, K. (2018).

Audiodateien[66] und Textabschnitte[67] wurden bereits Adversarial Examples konstruiert.

Adversarial Examples sind häufig übertragbar zwischen Modellen. Dies ist besonders im Zusammenhang mit Black-Box-Angriffen von Interesse. Hier bietet sich für eine:n Angreifer:in die Möglichkeit, anhand eines Ersatzmodells Adversarial Examples zu entwickeln, um diese dann beim Angriffsziel erfolgreich einzusetzen.

Verteidigungsmaßnahmen

Viele der für Data Poisoning diskutierten Verteidigungsmaßnahmen zielen darauf ab, »vergiftete« Beispiele zu identifizieren und von der Trainingsphase auszuschließen (Data Sanitization). Es existieren aber auch Verteidigungsmethoden für Modelle, die aufgrund erfolgreicher Backdoor Attacks bereits »vergiftet« sind. In diesem Fall konzentrieren sich die Verteidigungsmaßnahmen darauf, Muster, die als Auslöser für ein durch die Angreifenden erwünschtes Verhalten dienen sollen, während der Anwendungsphase aufzuspüren und nicht weiterzuverarbeiten.

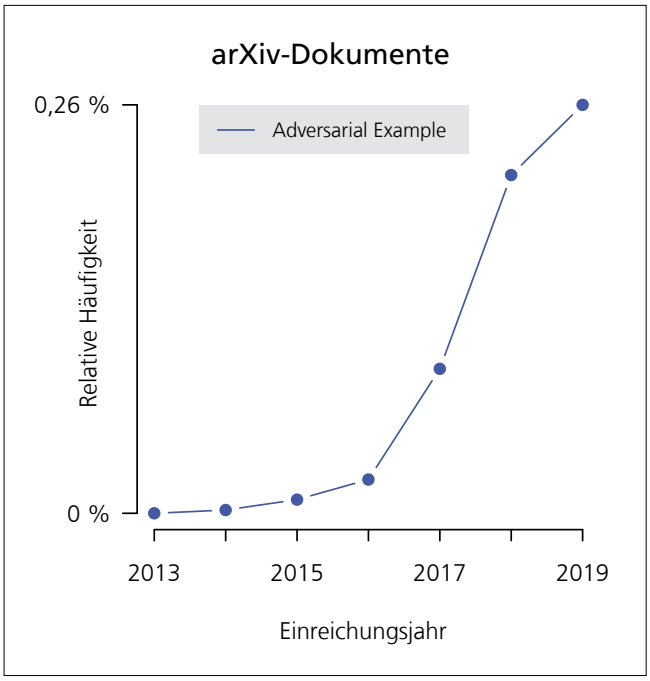


Abb. 28: Trendverlauf für Adversarial Examples nach Vorkommen im Titel oder Abstract. Erhebung vom 11.11.2019.

Verteidigungen gegen Angriffe mit Adversarial Examples streben an, die relative Häufigkeit der erfolgreichen Angriffe zu reduzieren. Idealerweise existieren dann keine Adversarial Examples mehr für ein Machine-Learning-Modell. Doch auch wenn die Erstellung von Adversarial Examples lediglich erschwert wird, kann dies aufgrund des Mehraufwands abschreckend auf Angreifende wirken. Die Verteidigungsmaßnahmen lassen sich dabei in zwei Typen unterscheiden. Reaktive Verteidigungsmaßnahmen versuchen, eingegebene Adversarial Examples zu erkennen oder zu bereinigen. Bei proaktiven Verteidigungsmaßnahmen wird versucht, generell robustere Modelle durch Veränderungen des Trainings oder der Architektur zu entwickeln.

Im wissenschaftlichen Bereich ist derzeit ein Wettlauf zwischen Verteidigungs- und Angriffsmöglichkeiten zu beobachten. Bisher hat sich jedoch keine Verteidigungsmaßnahme dauerhaft

[66] Carlini, N.; Wagner, D. (2018)

[67] Wang, W.; Wang, L.; Tang, B.; Wang, R.; Ye, A. (2019)

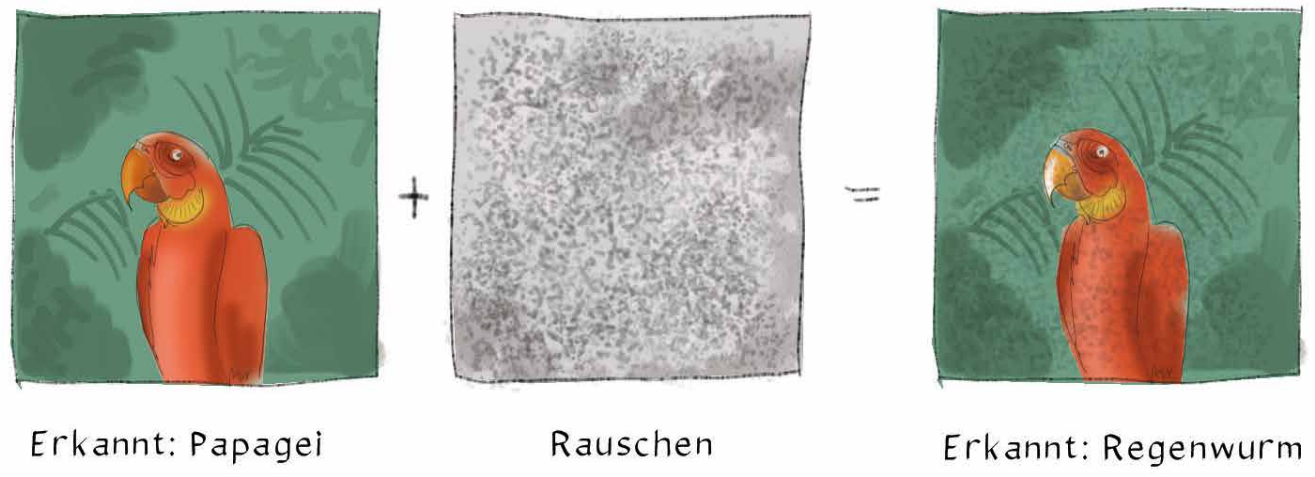


Abb. 29: Beispiel für ein durch Hinzufügen spezifische Rauschens erzeugtes Adversarial Example.

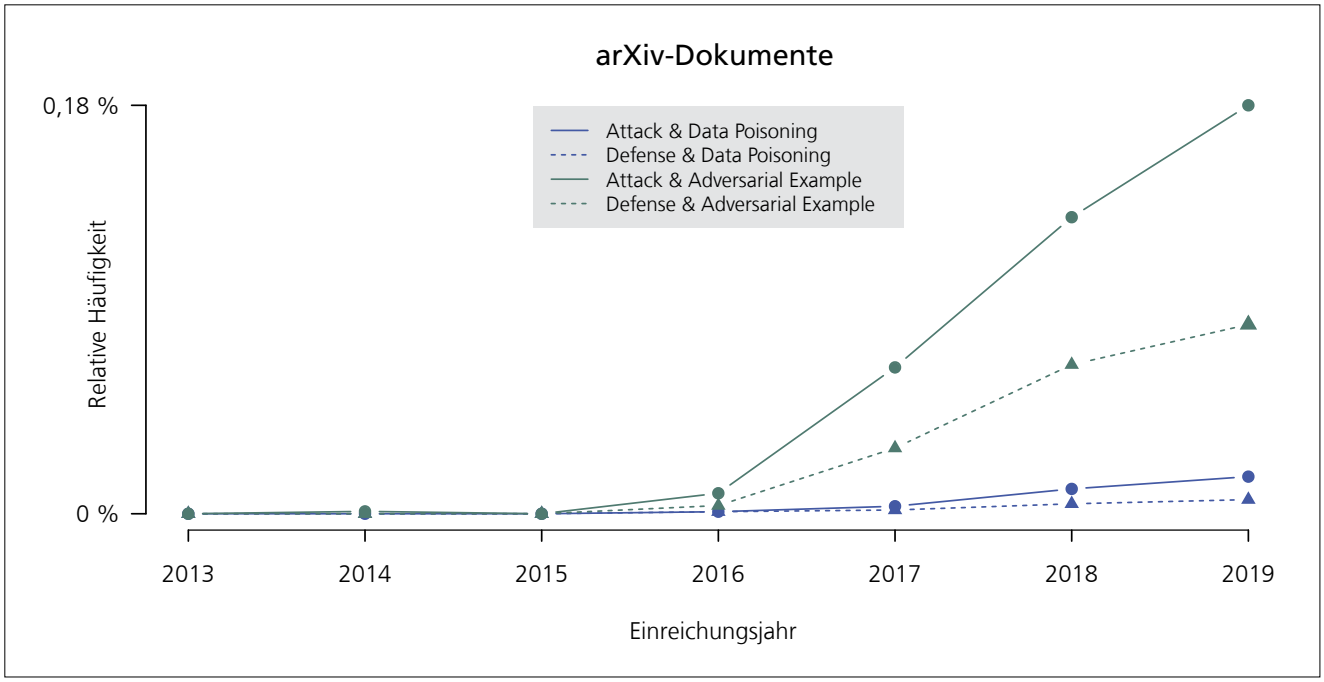


Abb. 30: Trendverlauf für Angriff & Verteidigung von KI nach Vorkommen im Titel oder Abstract. Erhebung vom 11.11.2019

bewährt. Tatsächlich existieren mittlerweile Forschungsergebnisse, die zu dem Schluss kommen, dass für die Robustheit gegenüber Adversarial Examples Obergrenzen existieren[68] und dass möglicherweise ein Zielkonflikt zwischen der Robustheit gegenüber Adversarial Examples und der allgemeinen Performanz von Modellen besteht[69].

Anwendungen

Poisoning Attacks und Adversarial Examples können dazu führen, dass Straßenschilder[70] oder sogar Fußgänger:innen[71] durch die Computer-Vision-Systeme autonomer Fahrzeuge übersehen oder mit anderen Zeichen, Verkehrsteilnehmer:innen oder Objekten verwechselt werden. Im Falle eines Stoppschilds genügte in einem der durchgeführten Experimente schon die Anbringung eines für Menschen unauffälligen Stickers, der gleichzeitig ein Adversarial Example darstellt. Eben aufgrund der Unauffälligkeit eines solchen Stickers wird möglicherweise erst etwas gegen diese Manipulation unternommen, nachdem sich bereits ein Unfall ereignet hat.

Auch automatische Gesichtserkennung lässt sich sowohl durch Adversarial Examples[72] als auch durch Data Poisoning[73] überlisten. Mögliche Anwendungen betreffen Privatpersonen (etwa bei der Entsperrung des Smartphones), Wirtschaft (etwa beim Zugang zum Unternehmensgelände) und den öffentlichen Sektor (etwa bei automatisierten Grenzkontrollen). Eine Täuschung der automatisierten Gesichtserkennung ist zum Beispiel schon mittels speziell für diesen Zweck angepasster Brillen möglich.

Weiterhin sind Sprachassistenten angreifbar[74]. Es ist möglich, Befehle an Sprachassistenten in Audiodateien so zu verstecken, dass Menschen sie gar nicht wahrnehmen. Ein Angreifer könnte beispielsweise Musik mit versteckten Befehlen an verschiedene Sprachassistenten frei im Internet zur Verfügung stellen. Dadurch ergibt sich die Möglichkeit, massenhaft Nutzer:innen auszuspionieren.

Architektur und Training eines künstlichen neuronalen Netzes können sehr aufwändig sein und erfordern mitunter enorme Kreativität. Ein gut funktionierendes neuronales Netz ist entsprechend wertvoll. Um das geistige Eigentum der Urheber:innen zu schützen, wurde daher vorgeschlagen, absichtlich Hintertü-

[68] Shafahi, A.; Huang, W. R.; Studer, C.; Feizi, S.; Goldstein, T. (2019).

[69] Tsipras, D.; Santurkar, S.; Engstrom, L.; Turner, A.; Madry, A. (2019).

[70] Siehe bspw. Akhtar, N.; Mian, A. (2018) und Gu, T.; Dolan-Gavitt, B.; Garg, S. (2017).

[71] Gu, T.; Dolan-Gavitt, B.; Garg, S. (2017).

[72] Kakizaki, K.; Yoshida, K. (2019).

[73] Chen, X.; Liu, C.; Li, B.; Lu, K.; Song, D. (2017).

[74] Siehe: https://nicholas.carlini.com/code/audio_adversarial_examples/.

ren einzubauen.[75] Eine solche Hintertür könnte so als Wasserzeichen dienen, ohne dass die Gesamtfunktionstüchtigkeit des neuronalen Netzes beeinträchtigt wird.

Weitere Bereiche, bei denen der Einsatz von Data Poisoning oder Adversarial Examples theoretisch denkbar ist oder bereits beobachtet wurde, sind die Erkennung von Schadsoftware[76], Empfehlungsdienste[77], Contentfilter[78] und die Auswertung medizinischer Bilder[79].

Auswirkungen, Chancen und Risiken

Die Verwundbarkeit von Machine-Learning-Modellen wirft grundsätzliche Fragen bezüglich der Anwendung dieser Modelle auf, insbesondere wenn es sich um sicherheitskritische Bereiche wie etwa den Straßenverkehr oder Passkontrollen handelt. Auch von Maschinen kann nicht erwartet werden, stets die korrekte Entscheidung zu treffen. Je nach Anwendungsbereich muss entschieden werden, wie viel Restrisiko inkorrekter Entscheidungen akzeptabel ist. Wenn keine dauerhaft effektiven Verteidigungsmaßnahmen entwickelt werden und Angriffe mit verhältnismäßig wenig Aufwand verbunden sind, könnte dies sogar prohibitiv für gewisse Einsatzbereiche wirken. Gleichzeitig bieten sich aber auch Chancen. Die Gefahr von Poisoning Attacks könnte zu einem sorgsameren Umgang mit der Datengrundlage und der Modellarchitektur führen, was zwar mehr Aufwand bedeutet, allerdings ultimativ zu besseren Modellen führen könnte. Zusätzliche Sicherheitsmaßnahmen wie etwa die parallele Verwendung mehrerer Machine-Learning-Verfahren unterschiedlicher Architektur mit »Mehrheitsentscheid« könnten die Gesamtverlässlichkeit von Systemen verbessern. Adversarial Examples könnten zudem helfen, künstliche neuronale Netze besser zu verstehen und so einen Beitrag zur Erklärbarkeit von KI leisten. Genauso wie die KI selbst, kann auch deren Verwundbarkeit letztlich situativ positiv oder negativ sein. So könnten sich für Privatpersonen neue Möglichkeiten bieten, sich den staatlichen beziehungsweise privatwirtschaftlichen Kontroll- und Überwachungsmaßnahmen zu entziehen, was je nach Person und Staat bzw. Unternehmen positiv oder negativ sein kann.

Handlungsempfehlungen

Forschung zur Messung der Verwundbarkeit von KI-Systemen fördern. Zwar existieren Ansätze, um die Robustheit von KI gegenüber Adversarial Examples und Data Poisoning messbar zu machen,[80] allerdings besteht hier weiterhin Forschungsbedarf.

Die Verwundbarkeit von KI-Systemen berücksichtigen. Neben klassischen Faktoren wie Gesamtkosten des Betriebs und Performanz sollte auch die Verwundbarkeit von KI-Systemen bei Entscheidungen über deren Einsatz berücksichtigt werden. Hierbei sollte besonders auf längerfristige Sicherheit geachtet werden.

Trainingsdaten und Modellarchitektur sorgsam auswählen. In der Entwicklung ist besonderes Augenmerk auf Datensicherheit und Daten-Qualitätssicherung zu legen, insbesondere, wenn Daten aus öffentlich verfügbaren Quellen in die Modellbildung eingehen. Auch die Modellarchitektur sollte sorgsam ausgewählt werden. Diese Maßnahmen sind speziell bei sicherheitskritischen Anwendungen von Bedeutung und der damit einhergehende Mehraufwand sollte nicht gescheut werden.

Die Vielfalt von KI-Technologien fördern. Auch KI-Technologien, die nicht auf neuronalen Netzen beruhen, sollten gefördert werden. Während neuronale Netze derzeit große Erfolge feiern, könnten sich andere Technologien langfristig als robuster erweisen und Monokulturen vermieden werden, bei denen einheitliche Angriffe größeren Schaden ausrichten können.

Flankierende Sicherheitsmaßnahmen auf Systemebene erforschen und fördern. Eine robuste IT-Infrastruktur kann Data Poisoning verhindern. Redundanz durch parallelen Einsatz unterschiedlicher Machine-Learning-Modelle könnte eine Möglichkeit darstellen, erfolgreiche Angriffe zu erschweren.

Erklärbare KI[81] fördern. Machine-Learning-Modelle, die ihre Entscheidung erklären, könnten bei der Ursachenforschung für Adversarial Examples helfen und Hintertüren enttarnen[82]. In der Fachliteratur wird zudem davon ausgegangen, dass Erklärbare KI robuster gegenüber Angriffen ist.[83] Hier besteht Forschungsbedarf. Darüber hinaus kann Erklärbare KI generell vertrauensschaffend wirken und so zur Akzeptanz von KI beitragen.

[75] Adi, Y.; Baum, C.; Cisse, M.; Pinkas, B.; Keshet, J. (2018).

[76] Paudice, A.; Muñoz-González, L.; Gyorgy, A.; Lupu, E. C. (2018).

[77] Fang, M.; Yang, G.; Gong, N. Z.; Liu, J. (2018).

[78] Olson, P. (2019).

[79] Ma, X.; Niu, Y.; Gu, L.; Wang, Y.; Zhao, Y.; Bailey, J.; Lu, F. (2019).

[80] Agarwal, C.; Dong, B.; Schonfeld, D.; Hoogs, A. (2018).

[81] Gahntz, M. (2019).

[82] Bachl, M.; Hartl, A.; Fabini, J.; Zseby, T. (2019).

[83] Fidel, G.; Bitton, R.; Shabtai, A. (2019).

Geografische Verortung

Die US-basierte Forschung ist bezüglich der Indikatoren für Qualität, Quantität und internationale Vernetzung mit Abstand führend im Bereich der Angreifbarkeit von KI. In Europa (Italien, Deutschland, Zypern, Luxemburg) wurde schon früh zum Innovationsfeld geforscht. Deutschland gehört bei Quantität und Qualität jeweils zu den besten 5 Staaten und ist am stärksten mit den USA vernetzt.

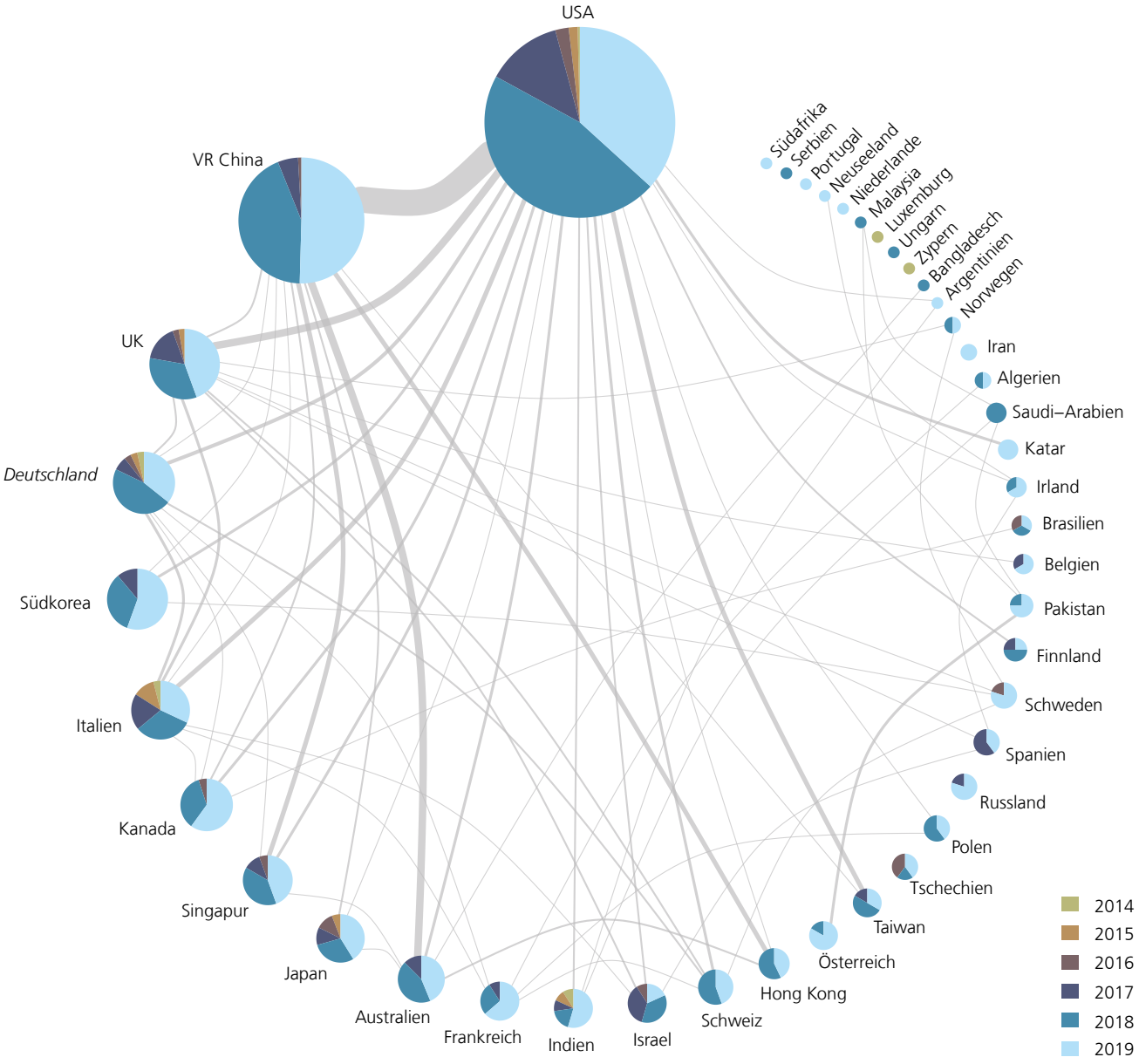


Abb. 31: Geografische Verteilung der Publikationshäufigkeit für das Innovationsfeld »Die Achillesferse der KI?«. Publikationsanzahl: 1 bis 264. Gemeinsame Publikationsanzahl: 1 bis 28. Erhebung vom 11.11.2019.

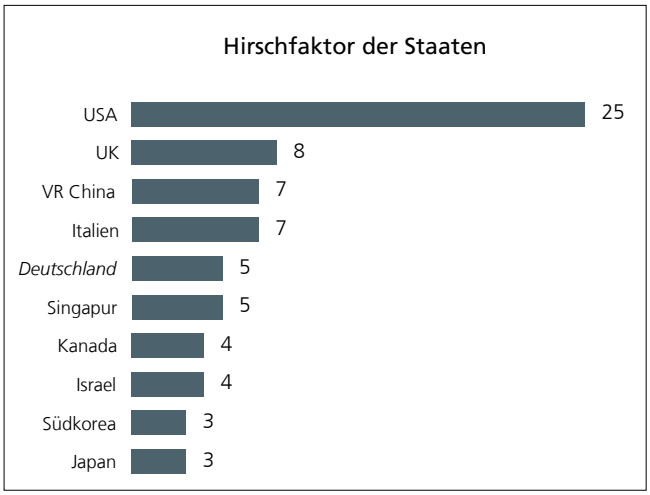


Abb. 32: Geografische Verteilung des Hirschfaktors für das Innovationsfeld »Die Achillesferse der KI?«. Erhebung vom 11.11.2019.

Wahrnehmung in Wissenschaft und sozialen Medien

Sowohl in der Wissenschaft, als auch in den sozialen Medien fokussiert sich die Diskussion stark auf künstliche Intelligenz und Sicherheit. Als Schlüsselwörter in den wissenschaftlichen Publikationen treten häufig verschiedene Angriffsarten und Machine-Learning-Architekturen auf. In den sozialen Medien werden auffällig häufig Programmiersprachen (Java, Python etc.) erwähnt. Bemerkenswert ist auch der Hashtag »#arxiv«, der auf die Diskussion neuer wissenschaftlicher Ergebnisse in den sozialen Medien hinweist.

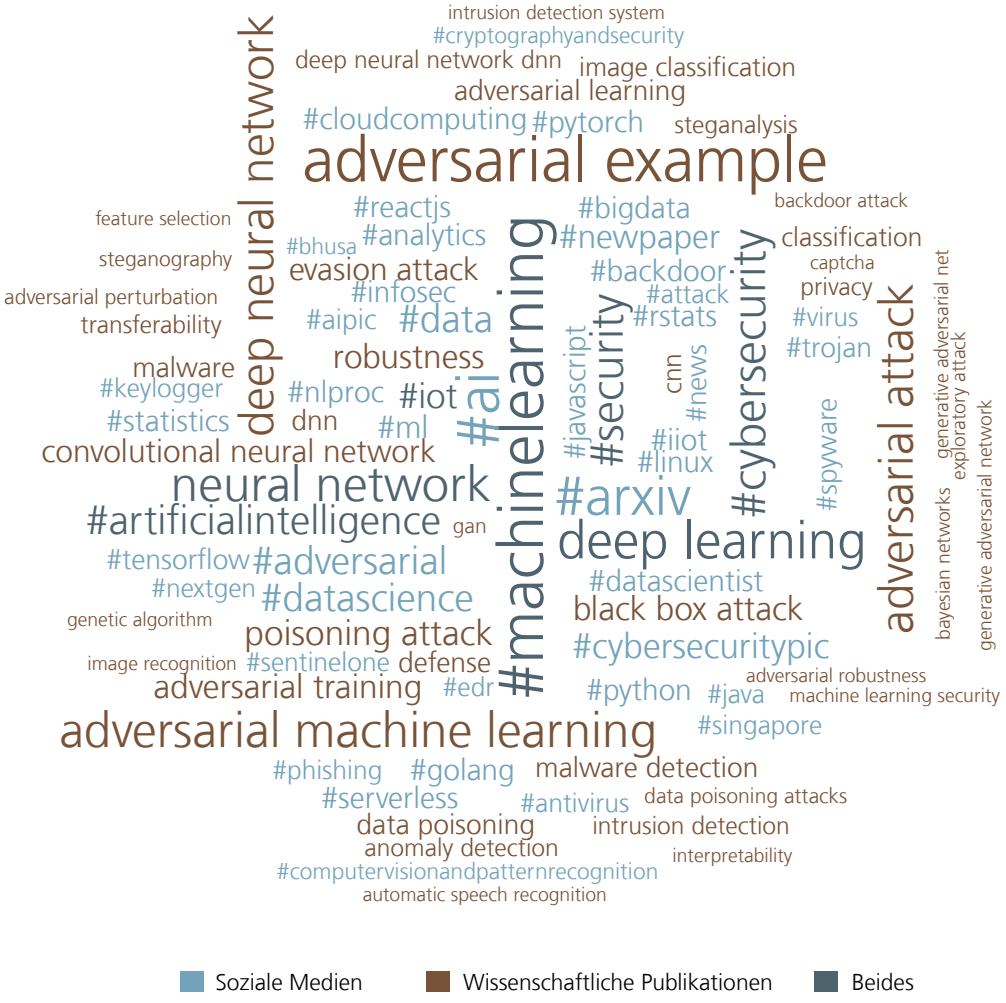


Abb. 33: Wordcloud für das Innovationsfeld »Die Achillesferse der KI?«. Die häufigsten Hashtags und Schlüsselwörter für die Jahre 2018 und 2019. Erhebung vom 11./12.11.2019.

2.5 MASCHINEN VERSTEHEN MENSCHEN

Je enger der Kontakt von Mensch und Maschine, desto wichtiger wird es, dass Maschinen Menschen verstehen und menschliche Absichten und menschliches Verhalten korrekt interpretieren. Dabei geht es um die Echtzeitanalyse von Text und Sprache, Mimik, Körpersprache und Bewegungen ebenso wie z. B. um die Erkennung und Unterscheidung von verbaler Gewalt und Satire in Onlineforen und sozialen Medien. Mit steigender Rechenleistung und Möglichkeiten der automatisierten Auswertung großer Datenmengen entstehen im Bereich des maschinellen Lernens neue Techniken, um solche Auswertungen vorzunehmen bzw. andere Ansätze zu ergänzen. Noch sind alle Ansätze mit einer Reihe von Unzulänglichkeiten behaftet und können durch vergleichsweise einfache Variationen getäuscht werden. Mit weiteren Fortschritten im Bereich des Natural Language Processing, der Computer Vision, Predictive Analytics und Anomalieerkennung sowie der Kombination unterschiedlicher Ansätze können Aussagen, Intentionen und Verhalten jedoch immer präziser verstanden werden.

Hate Speech Detection

2016 wurden jede Minute 3,3 Mio. Facebook-Posts erstellt, 450.000 Tweets abgesetzt, 500 Stunden Videomaterial auf YouTube hochgeladen und 66.000 Fotos bei Instagram geteilt.[84]

[84] Allen, R. (2017).

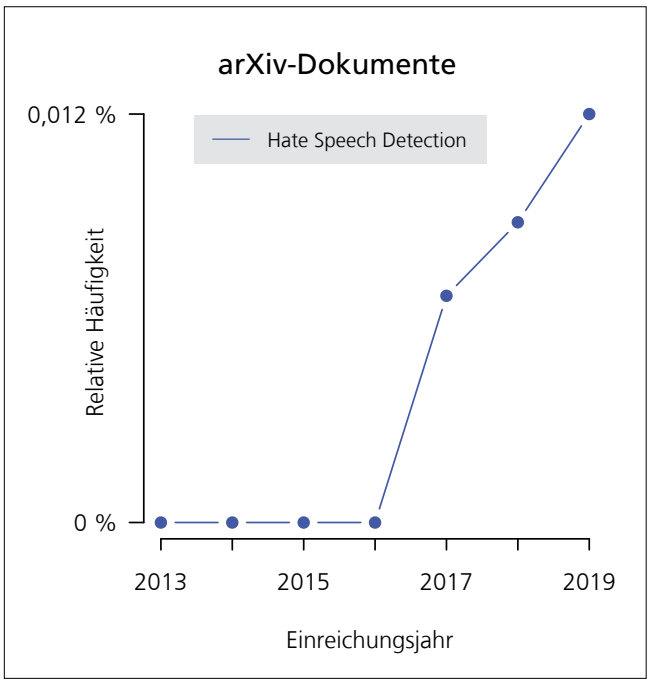


Abb. 34: Trendverlauf für Hate Speech Detection nach Vorkommen im Titel oder Abstract. Erhebung vom 04.11.2019

Diese Menge an Nutzer:innen-generierten Inhalten macht es schier unmöglich, allein mithilfe menschlicher Content-Moderator:innen Hassrede, Beleidigungen, Gewaltdarstellungen, terroristische Propaganda und strafrechtlich relevante Inhalte kurzfristig zu erkennen und zu entfernen. Mit der großen Verbreitung sozialer Medien wächst auch die Anzahl derer, die diese für ihre Zwecke missbrauchen. Dies kann das Nutzer:innenerlebnis schmälern und sich geschäftsschädigend auswirken.[85] Die automatisierte Erkennung von Hassrede (Hate Speech Detection) wird daher immer bedeutsamer. Hassrede und strafrechtlich relevante Inhalte sind hierbei von bloßer derber Sprache zu unterscheiden.[86] Eine besondere Schwierigkeit ist, dass das Löschen von Posts und Inhalten direkt in das Grundrecht der Meinungsäußerungsfreiheit eingreift. Hier gilt es daher sicherzustellen, dass tatsächlich nur Inhalte entfernt werden, die gegen geltendes Recht oder die jeweiligen Community-Standards verstoßen, wobei zwischen beiden auch Widersprüche bestehen können.

Es gibt unterschiedliche Vorgehensweisen, um automatisiert Hassrede in Texten zu erkennen. Hierfür wird ein Text zunächst in einzelne Wörter oder Ausdrücke bzw. Konzepte unterteilt. Das einfachste Vorgehen ist der Abgleich der enthaltenen Wörter mit Sentiment- und Emotionslexika bzw. einer Blacklist Hassrede-indizierender Begriffe. Dieses Vorgehen birgt jedoch mehrere Schwierigkeiten: Zum einen können indizierte Begriffe durch eine veränderte Schreibweise abgewandelt werden und hierdurch einer automatisierten Detektion entgehen.[87] Zum anderen ist Hassrede kontextabhängig, sodass indizierte Begriffe nicht notwendigerweise nur bei Hassrede verwendet werden (sondern bspw. auch ironisch) und gleichzeitig Hassrede auch durch für sich genommen harmlose Wörter ausgedrückt werden kann.[88] Zudem ist die Erstellung entsprechender Lexika sehr aufwändig, weil sich Sprache stetig weiterentwickelt.[89]

Elaboriertere Verfahren greifen auf Methoden aus dem Natural Language Processing (NLP) zurück. Hierbei werden Wörter und Ausdrücke bzw. Konzepte als Vektoren dargestellt, über die ihre semantische, grammatikalische und strukturelle Nähe zu bereits bekannten (in der hier betrachteten Anwendung: Hassrede beinhaltenden) Textstücken erfasst werden kann. Hierdurch können auch scheinbar neutrale Begriffe als eher positiv oder negativ klassifiziert werden.

[85] Djuric, N.; Zhou, j.; Morris, R.; Grbovic, M.; Radosavljevic, V.; Bhamidipati, N. (2015).

[86] Davidson, T.; Warmsley, D.; Macy, M.; Weber, I. (2017).

[87] siehe Fussnote 85.

[88] Nobata, C.; Tetreault, J.; Thomas, A.; Mehdad, Y.; Chang, Y. (2016).

[89] Davidson, T.; Warmsley, D.; Macy, M.; Weber, I. (2017).

Für die Repräsentation der Wörter, Wortgruppen oder Texte als Vektoren gibt es eine Reihe unterschiedlicher Verfahren. Basierend auf Vektoren für Beispieldaten bereits bekannter Einordnung werden Klassifikatoren trainiert. Nachdem alle Wörter auch in ihrem Zusammenhang klassifiziert worden sind, treffen Hate-Speech-Detection Algorithmen die Entscheidung, ob ein gegebener, vorher unbekannter Text als Hassrede einzustufen ist.[90] Neuere Ansätze experimentieren auch mit tiefen neuronalen Netzen, die mithilfe verschiedener Architekturen spezifisch darauf trainiert werden, semantische Einordnungen in Bezug auf Hassrede vorzunehmen. Die Techniken können auf unterschiedliche Weise modifiziert und kombiniert werden.[91]

Die Genauigkeit der automatisierten Erkennung von Hassrede wird üblicherweise über einen Vergleich mit von Menschen generierten Bewertungen erfasst. Auch letztere kann jedoch uneindeutig sein und verschiedenen Fehlern unterliegen.[92]

Sarkasmus- und Ironieerkennung

Während die Erkennung von Hassrede im Internet ein wichtiges Anwendungsfeld darstellt, gilt es dabei zugleich die Meinungsäußerungsfreiheit zu gewährleisten und keine Inhalte zu löschen, die weder strafrechtlich relevant sind noch gegen die Community-Standards verstoßen. Besonders problematisch ist diese Unterscheidung bei ironischen oder satirischen Beiträgen, die sich ähnlicher (Bild-)Sprache bedienen wie Hassrede, diese aber übertreiben und dadurch in ihr Gegenteil verkehren. Fällt Menschen diese Unterscheidung schon häufig schwer, stellt dies Maschinen vor noch größere Herausforderungen. Die hierfür angewandten Techniken sind in vielen Fällen identisch mit jenen für die Erkennung von Hassrede.[93] Syntax, Affekt-Ladung, Kontext, Linguistik, Semantik, diskursive Marker und zusätzliche Informationen, bspw. über den/die Autor:in, spielen jedoch nach gängiger Meinung eine größere Rolle[94], obwohl diese – sofern möglich – auch bei der Erkennung von Hassrede genutzt werden sollten, um Falschalarme zu minimieren. Dabei wird davon ausgegangen, dass es kein eindeutiges Unterscheidungsmerkmal gibt, um Ironie zu erkennen, sondern verschiedene Merkmale in ihrer Gesamtheit Aufschluss geben können, wobei manche Merkmale bessere Indikatoren sind als andere.[95] Auf

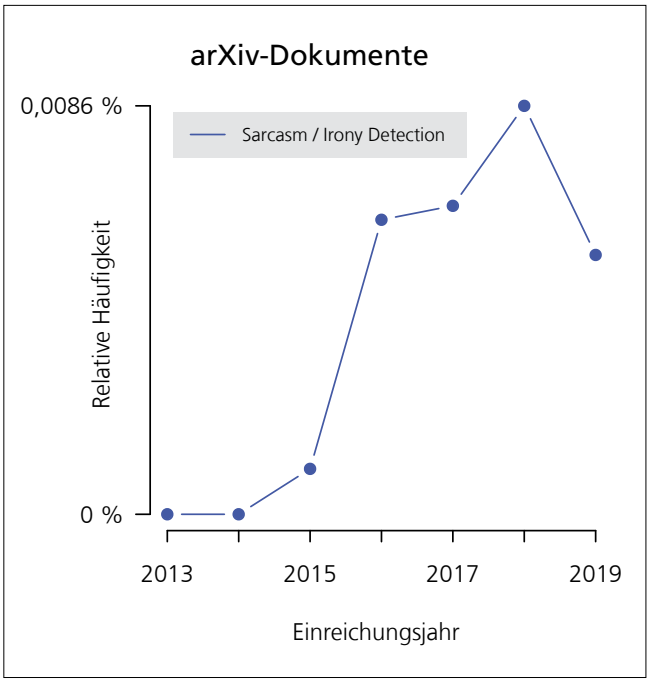


Abb. 35: Trendverlauf für Sarkasmus- und Ironieerkennung nach Vorkommen im Titel oder Abstract. Erhebung vom 05.11.2019

Syntaxebene können bspw. Satzzeichen, Emoticons, Sonderzeichen, URLs, Hashtags, Begriffe, Wortarten, Muster, Groß- und Kleinschreibung, Wort- und Textlänge und Worthäufigkeit Aufschluss über die Intention des/der Autor:in geben.[96]

Da ein ironischer Effekt auch dadurch erzeugt wird, dass eine Inkongruenz zwischen der erwarteten und der tatsächlichen Aussage besteht, gibt es ebenfalls Ansätze, die versuchen, solche Disbalancen aufzudecken und hierüber Ironie zu erkennen. Je stärker bspw. positiv und negativ konnotierte Hinweisgeber (wie Wörter, aber bspw. auch Emoticons) in direkter Nachbarschaft auftauchen, desto höher die Wahrscheinlichkeit für Ironie. Die Inkongruenz kann sich aber auch extern in der Beziehung zum Textgegenstand, der Situation der Äußerung oder dem/der Autor:in ergeben.[97]

Auch über die Unterscheidung von Hassrede hinaus gibt es verschiedene Einsatzgebiete für Ironie- und Sarkasmuserkennung, bspw. die Sentimenterkennung bei Produktbewertungen, Opinion Mining, Social-Media-Analyse oder Empfehlungs- und Dialogsysteme. Auch hier können automatisierte Systeme in die Irre geführt werden, wenn sie ironische Aussagen nicht als solche verstehen, sondern für bare Münze nehmen.[98] Da der ironi-

[90] Davidson, T.; Warmsley, D.; Macy, M.; Weber, I. (2017).

[91] Badjatiya, P.; Gupta, S.; Gupta, M.; Varma, V. (2017).

[92] Davidson, T.; Warmsley, D.; Macy, M.; Weber, I. (2017).

[93] Siehe bspw. Hernandez-Farias, D. I.; Patti, V.; Rosso, P. (2016) und Zhang, S.; Zhang, X.; Chan, J.; Rosso, P. (2018).

[94] Siehe bspw. Wallace, B. C. (2013) und Hernandez-Farias, D. I.; Patti, V.; Rosso, P. (2016).

[95] Reyes, A.; Rosso, P. (2013).

[96] Nobata, C.; Tetreault, J.; Thomas, A.; Mehdad, Y.; Chang, Y. (2016).

[97] Hernandez-Farias, D. I.; Patti, V.; Rosso, P. (2016).

[98] Reyes, A.; Rosso, P. (2013).

sche Effekt häufig darüber erreicht wird, dass negative Gefühle durch eine übertrieben positive Wortwahl ausgedrückt werden (oder in Ausnahmefällen auch umgekehrt), werden in diesem Fall Meinungen falsch klassifiziert.[99]

Problematisch ist bei der Ironieerkennung bereits die Erstellung von Trainingsdaten. Neben Texten, die von den Autor:innen selbst als Ironie oder Sarkasmus tituliert werden (bspw. über Hashtags), kommen auch von Dritten gelabelte Trainingsdaten zum Einsatz. Hierbei zeigt sich, dass auch Dritte – insbesondere bei einzelnen Sätzen ohne Kontext – über das Vorhandensein von Ironie häufig uneins sind.[100]

Intent Recognition

Bei der Intent Recognition geht es darum, Intentionen oder menschliches Verhalten möglichst präzise zu erkennen und dadurch – wenn möglich – sogar vorherzusagen. Mit der Zunahme von Mensch-Maschine-Interaktionen (z. B. in der vernetzten Produktion, der Smart City oder beim automatisierten Fahren) kommt der maschinellen Intentionserkennung eine immer wichtigere Rolle zu, um bspw. Unfälle zu vermeiden. Aber auch in der Rehabilitation und bei Sprachassistenten ist es erforderlich, Absichten zu erkennen und darauf zu reagieren. Hierfür werden bspw. Körper-, Text- und Sprachdaten analysiert (siehe Abb. 37). Für die Überwachung des öffentlichen Raums mit intelligenten Kameras sind insbesondere die Echtzeiterkennung und -auswertung von Bewegungsmustern von Interesse. Damit soll bestimmt werden, ob eine Situation eine Gefahr für die öffentliche Sicherheit darstellt oder sich zu einer solchen entwickeln könnte, wie Unfall, Diebstahl oder Schlägerei. Hierbei geht es häufig um die Anomalieerkennung und -vorhersage bei Überwachungsvideos, für die neuronale Netze mit nicht überwachten Lernmethoden trainiert werden. Die Herausforderung ist, potenziell gefährliche Situationen möglichst zuverlässig zu erkennen und dabei möglichst wenige Falschalarme zu erzeugen.

Bei Sparse-Coding-Ansätzen wird aus einem kleinen Teil des initialen Videomaterials eine Datenbank erstellt, die nur normale Ereignisse umfasst. Irreguläre Ereignisse lassen sich mit dieser Datenbank nicht abbilden und werden daher als Anomalie erkannt. Neue Ansätze nutzen auch irreguläre Ereignisse für das Training, weil die Unterschiede häufig fließend oder kontextabhängig sind. So gibt es viele Ereignisse, die zwar von dem initialen Trainingsmaterial abweichen, aber trotzdem keine

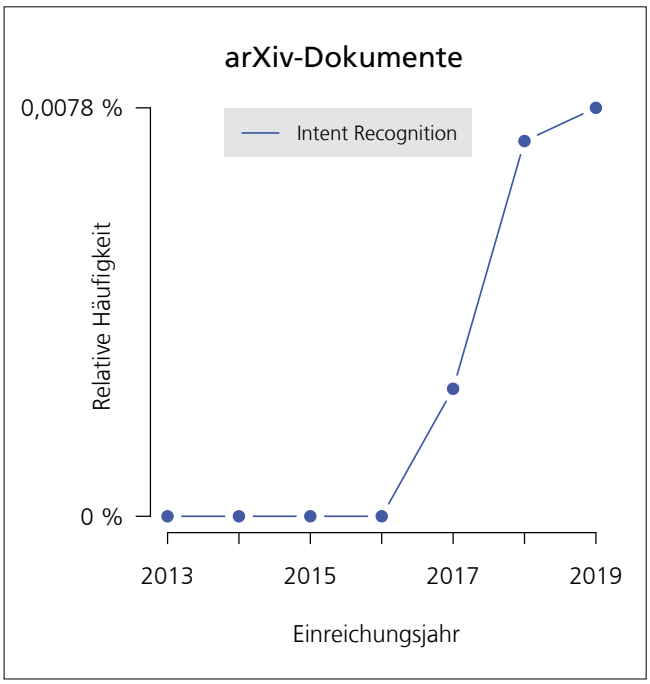


Abb. 36: Trendverlauf für Intent Recognition nach Vorkommen im Titel oder Abstract. Erhebung vom 04.11.2019

Gefahr darstellen, bspw. Veränderungen je nach Jahres- oder Tageszeit. Gleichzeitig kann ein und dasselbe Verhalten in einem Fall eine Gefahr darstellen, in einem anderen jedoch nicht. Schlechte Sichtverhältnisse oder das Verdecken von Objekten stellt die Videoanalyse ebenfalls noch vor Herausforderungen.[101]

Auch bei Sprachassistenten kommt der Intentionserkennung eine wesentliche Rolle zu. Hierbei wird häufig mehrstufig vorgegangen: Zuerst wird erfasst, auf welchen Bereich sich eine Anfrage oder ein Befehl bezieht. Dann werden die sinntragenden Wörter und ihre Beziehung zueinander identifiziert (bspw. Namen, Orte oder Zeiten und entsprechende Verben), um die Intention zu erfassen.[102]

Bei der maschinellen Unterstützung von Bewegungen, bspw. beim Gehen, ist es wichtig, Bewegungen zu erkennen und möglichst präzise vorherzusagen. Hierfür werden Körpersensoren genutzt, die bspw. die Bewegung von Muskeln und Gelenken erfassen. Einige Ansätze bauen aus diesen Daten nach und nach ein systeminternes Erfahrungswissen auf, das gemeinsam mit den Echtzeitdaten die Vorhersagefähigkeiten verbessern kann.[103]

[99] Hernandez-Farias, D. I.; Patti, V.; Rosso, P. (2016).

[100] Reyes, A.; Rosso, P. (2013); Hernandez-Farias, D. I.; Patti, V.; Rosso, P. (2016); Zhang, S.; Zhang, X.; Chan, J.; Rosso, P. (2018).

[101] Sultani, W.; Chen, C.; Shah, M. (2018).

[102] Hakkani-Tür, D.; Tur, G.; Celikyilmaz, A.; Chen, Y.; Gao, J.; Deng, L.; Wang, Y. (2016).

[103] Martinez-Hernandez, U.; Dehghani-Sanij, A. A. (2018).

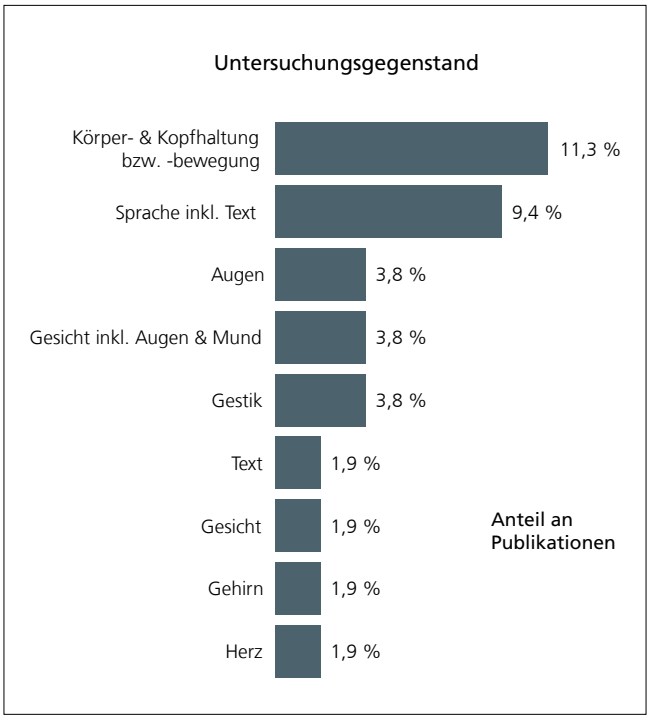


Abb. 37: Analyse des Untersuchungsgegenstands[104] für Intent Recognition anhand von Scopus-Daten für die Jahre 2013 bis 2019. Erhebung vom 04.11.2019.

Anwendungen

Intentionserkennung spielt eine wichtige Rolle in der Mensch-Maschine-Interaktion, bspw. in Smart Cities, beim autonomen Fahren, in der Industrie 4.0, im Bereich der Telemedizin und Rehabilitation oder im Smart Home, aber auch für die öffentliche Sicherheit, Chatbots und Sprachassistenten und in den Bereichen Kundenservice, Marketing und Sales.

Die treffsichere automatisierte Erkennung und Unterscheidung von Hassrede und Ironie spielt mit dem steten Wachstum an Nutzer:innen-generierten Inhalten eine immer wichtigere Rolle in der Content-Moderation in sozialen Medien, Internetforen und Kommentarbereichen. Sie kann einen Beitrag leisten zum Kinder- und Jugendschutz, zur Verbesserung des Nutzer:innenerlebnisses, zur Wahrung von Geschäftsinteressen der Websitebetreiber:innen und Werbetreibenden, zur Strafverfolgung und zur Terrorismusbekämpfung. Zudem liefert sie interessante Erkenntnisse für die Entwicklung von Chatbots und Sprachassistenten sowie Meinungsforschung und Psychologie.

[104] Die Erkennung von Hassrede, Sarkasmus oder Ironie beruht hauptsächlich auf der Verarbeitung von Sprache. Bei Intent Recognition ist das anders, bspw. werden Gesichtsausdrücke und Herzrhythmus analysiert. Um einen Eindruck zu vermitteln, was vergleichsweise häufig als Untersuchungsgegenstand auftritt, wurden wiss. Publikationen auf das Vorkommen von Begriffen aus einer für diesen Zweck angelegten Stichwortsammlung überprüft.

Auswirkungen, Chancen und Risiken

Je treffgenauer Maschinen die Inhalte und Intentionen menschlicher Kommunikation und menschlichen Verhaltens einschätzen können, desto reibungsloser funktioniert die Mensch-Maschine-Interaktion und desto besser kann die Meinungsfreiheit im digitalen Raum geschützt werden. Die herausfordernde Arbeit der Identifikation (möglicherweise) menschenverachtender Beiträge oder der Beobachtung von Kamerafeeds könnte in diesem Fall von Maschinen übernommen werden und zu erheblichen Effizienzgewinnen führen.

Durch die automatisierte Erkennung von Inhalten und Intentionen sollen sichere Räume geschaffen werden, im Analogen wie im Digitalen. Doch selbst wenn der Hass im Netz unsichtbar gemacht wird, ist fraglich, ob er dadurch tatsächlich abnimmt. Zumindest könnte verhindert werden, dass sich Gleichgesinnte in ihren extremen Ansichten weiter bestärken. Diese werden im Zweifel jedoch auf andere Kommunikationsnetzwerke ausweichen, wodurch sich Filterblasen sogar noch verstärken könnten.

Wichtig ist in diesem Zusammenhang auch die Frage, wer die Techniken entwickelt und wer sie auf welche Weise einsetzt. So weichen die Vorstellungen davon, was Hassrede darstellt und was von der Meinungsfreiheit gedeckt ist, in verschiedenen Ländern sowie zwischen Staaten und den Community-Standards internationaler Konzerne deutlich voneinander ab. Zudem treiben auch nicht-demokratische Staaten die Forschung in diesem Bereich entschieden voran. Da die entwickelten Sprach-erkennungstechniken nicht nur auf Hassrede und Ironie anwendbar sind, können sie ebenso für die Unterdrückung oppositioneller oder regimekritischer Inhalte und zum Aufbau von Überwachungsstrukturen genutzt werden.

Es kann auch problematisch sein, wenn illegitime Inhalte automatisch gelöscht werden, ohne dass eine Beweissicherung stattfindet oder mit Strafverfolgungsbehörden kooperiert wird. Hierdurch können frühzeitige Hinweise auf gewaltbereite Täter:innen verloren gehen und auch die Aufklärung von Straftaten oder die effektive Strafverfolgung könnte vereitelt werden. Wenn strafrechtlich relevante Inhalte online gar nicht erst geteilt werden können, weil diese automatisiert erkannt und blockiert werden, stellt dies das Prinzip der Gesetzestreue auf den Kopf. Was bedeutet es für die Akzeptanz von Gesetzen, wenn Verstöße technisch nicht mehr möglich sind? Inwiefern werden mit ausreichend Motivation und Ressourcen versehene Nutzer:innengruppen alle Möglichkeiten nutzen, die hier erörterten Schutzmechanismen zu unterlaufen (siehe Kapitel 2.4) und mittels eines fortgesetzten »Wettrüstens« eine nicht vernachlässigbare fortdauernde Präsenz ihrer Inhalte erzielen?

Bei Design und Einsatz automatisierter Inhalts- und Verhaltensanalysetechniken ist zudem darauf zu achten, dass diese die Diskriminierung marginalisierter Gruppen nicht systematisieren. So werden zur Berechnung der Erfolgsquote häufig die maschinellen Ergebnisse mit den Eingaben von Menschen abgeglichen. Dabei ist jedoch oft nicht nachzuvollziehen, ob die menschlichen Eingaben valide und vorurteilsfrei sind. Auch bei der Verhaltensanalyse könnten Menschen immer wieder in den Fokus geraten, weil sie sich bspw. aufgrund einer Behinderung anders bewegen.

Insgesamt ermöglichen die hier vorgestellten Techniken Erkenntnisgewinne darüber, wie wir kommunizieren, denken und handeln. Da die Masse an Nutzer:innen-generierten Inhalten und an Überwachungskameras im öffentlichen Raum beständig wächst, ist es bereits jetzt nicht mehr möglich, die Daten schnell und zuverlässig von Menschen auswerten zu lassen. Die Abhängigkeit von automatisierten Analysen wird in Zukunft noch weiter steigen. Aus diesem Grund sind die Validität und Zuverlässigkeit der Analysen sowie die Vorgehensweise, Anwendungs- und Geltungsbereiche und die betrachteten Parameter wichtige Fragen, die über die Zulässigkeit des Einsatzes entscheiden sollten. Im Hinblick auf Validität und Zuverlässigkeit der Auswertungen stellt sich die Frage, wer diese im laufenden Betrieb überprüft und auf welche Art und Weise. Hierbei geht es vor allem auch darum, ein Overblocking legitimer Inhalte zu vermeiden und dadurch die Meinungsäußerungsfreiheit nicht zu gefährden. Da Sprache viele Nuancen kennt, wird zudem oft ein simplifiziertes Modell angewandt, das nicht alle Formen verbaler Gewalt gleich gut erkennt. Wenn zur Bewertung der Inhalte ergänzende Informationen genutzt werden, kann dies jedoch wiederum zu einer Ungleichbehandlung führen, bei der ein und derselbe Inhalt von einer Nutzerin geteilt werden kann, bei einem anderen Nutzer jedoch blockiert wird.

Handlungsempfehlungen

Effektive Strafverfolgung ermöglichen. Während das Netzwerkdurchsetzungsgesetz Privatunternehmen in die Pflicht nimmt, Hassrede und strafrechtlich relevante Beiträge auf ihren Kommunikationsplattformen zu löschen, muss gleichzeitig sichergestellt werden, dass den Sicherheitsbehörden relevante Inhalte zur Ermittlung oder Strafverfolgung zur Verfügung gestellt werden. Die Sicherheitsbehörden sollten ebenso selbst über entsprechende Technik verfügen, um auch unabhängig von Privatunternehmen Inhalte ermitteln zu können.

Interdisziplinäre und inklusive Forschung fördern. Die beschriebenen Analysetechniken sind noch mit einer Reihe von Unzulänglichkeiten behaftet, die weitere Forschung erfordern. Weil Maschinen häufiger menschliche Aufgaben übernehmen und Entscheidungen treffen, sind neben technischer Expertise auch sektorspezifische und humanwissenschaftliche Expertisen einzubinden. Auch auf Diversität und die Einbindung Betroffener sollte geachtet werden. Zudem sollten die Belange marginalisierter Gruppen vertreten und Mehrsprachigkeit abbildbar sein.

Technische Möglichkeiten umsichtig nutzen. Die automatisierte Klassifizierung von Inhalten und Verhalten bietet Möglichkeiten zur Schaffung sicherer Räume, zur Gewährleistung der Meinungsfreiheit, zum Kinder- und Jugendschutz sowie zur Verfolgung von Straftaten. Diese Möglichkeiten sollten genutzt und mit anderen Ansätzen kombiniert werden. Dabei gilt es jedoch auch, das Gleichgewicht zu wahren zwischen Freiheit und Sicherheit. Angesichts der Möglichkeit, schädliches Verhalten durch technische Mittel erst gar nicht zuzulassen, stellt sich diese Frage immer wieder neu.

Missbrauch verhindern. Die vorgestellten Techniken können auch eingesetzt werden, um Autokratien stützen. Es gilt daher Wege aufzuzeigen, wie verhindert werden kann, dass diese Techniken missbräuchlich angewendet werden. Im Bereich der Wissenschaft sollte darauf geachtet werden, mit wem und zu welchen Zwecken geforscht wird und ob dabei ethische Standards eingehalten werden. Des Weiteren sollten die technischen Möglichkeiten der eingesetzten Systeme transparent nachvollziehbar sein, wo sie eingesetzt werden, sowie welche Daten hierbei gespeichert und verarbeitet werden. Falls der tatsächliche Einsatz über die gemachten Angaben hinausgeht, sollte dieser Verstoß entsprechend hart sanktioniert werden. Die Datenschutzgrundverordnung bietet hier entsprechende Sanktionsmechanismen. Zudem sollten sich Privatunternehmen die Option offenhalten, den weiteren Einsatz ihrer Techniken zu unterbinden, sofern diese vertragswidrig zu antidemokratischen Zwecken eingesetzt werden.

Nachweis der Eignung einfordern. Beim Einsatz von Techniken zur Erkennung von Hassrede, Ironie oder Verhalten passieren immer wieder Fehler: Beiträge werden zu Unrecht gelöscht oder Nutzer:innen gesperrt, gewaltverherrlichende oder extremistische Inhalte oder gefährliche Situationen nicht erkannt und Falschalarme ausgelöst. Dies kann für Einzelne gravierende Konsequenzen haben. Deshalb ist es unabdingbar, dass der Einsatz solcher Techniken von unabhängiger Seite regelmäßig überprüft wird. Dabei sollten einerseits die Validität und Zuverlässigkeit der Analyseergebnisse im Mittelpunkt stehen, aber auch die berücksichtigten Daten und die Fehlerquoten für unterschiedliche Gruppen. Das Ergebnis der Überprüfung sollte darüber entscheiden, ob die Technik für den jeweiligen Zweck weiter eingesetzt werden darf. Zudem sollten Wege zur Verfügung stehen, um Fehlklassifikationen schnell und einfach korrigieren zu lassen.

Geografische Verortung

Sowohl in Bezug auf die Anzahl der wissenschaftlichen Publikationen als auch mit Blick auf die wissenschaftliche Exzellenz sind die USA in der Forschung zum maschinellen Verständnis menschlicher Kommunikation und menschlichen Verhaltens klar führend. Auch in Indien und der VR China wird die Forschung erfolgreich vorangetrieben. In Europa dominieren Italien und Spanien die Forschung, die hierbei eng kooperieren. Deutschland liegt im Vergleich zu den anderen Innovationsfeldern etwas weiter zurück (Platz 10 beim Hirschfaktor und Platz 11 bei der Publikationshäufigkeit).

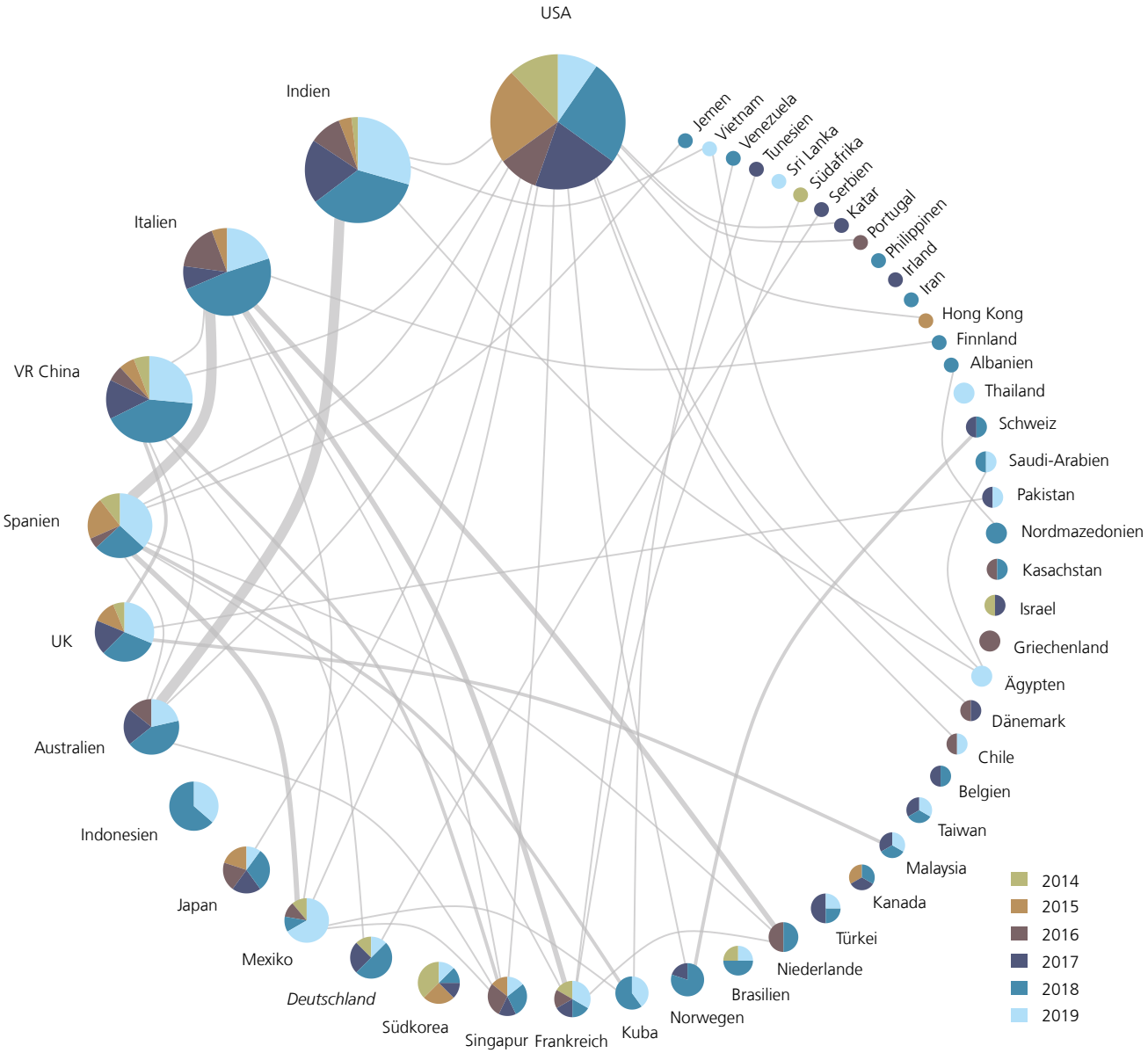


Abb. 38: Geografische Verteilung der Publikationshäufigkeit für das Innovationsfeld »Maschinen verstehen Menschen«. Publikationsanzahl: 1 bis 83. Gemeinsame Publikationsanzahl: 1 bis 6. Erhebung vom 04./05.11.2019.

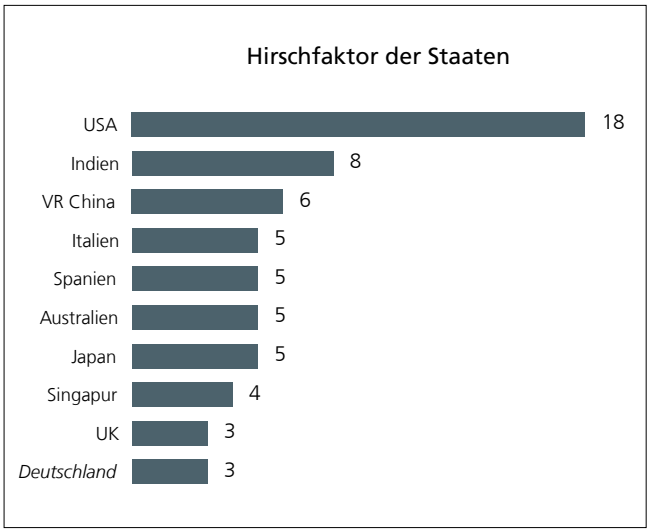


Abb. 39: Geografische Verteilung des Hirschfaktors für das Innovationsfeld »Maschinen verstehen Menschen«. Erhebung vom 04./05.11.2019.

Wahrnehmung in Wissenschaft und sozialen Medien

Die maschinelle Erkennung menschlicher Absichten und menschlichen Verhaltens ist äußerst vielfältig. Dies zeigt sich auch bei der Betrachtung der häufigsten Schlüsselwörter. Während in der Wissenschaft technikbezogene Stichworte – bspw. Namen spezifischer Arten neuronaler Netze – überwiegen, stehen in den sozialen Medien übergeordnete Schlagworte im Mittelpunkt, wie bspw. künstliche Intelligenz. Trendthemen wie die Erkennung von Hassrede und Sarkasmus oder Sentimentanalyse finden sich in beiden Medien.

Abb. 40: Wordcloud für das Innovationsfeld »Maschinen verstehen Menschen«. Die populärsten Hashtags und Schlüsselwörter für die Jahre 2018 und 2019. Erhebung vom 17.11.2019.

3. INNOVATIONSFELDER 2013

Zu den ersten Publikationen des Kompetenzzentrums Öffentliche IT gehörte 2013 die »ÖFIT-Trendschau: Innovationsfelder Öffentlicher IT« . Dabei handelte es sich sowohl um eine Übersicht quantitativ und qualitativ ermittelter Innovationsfelder als auch den Auftakt zur ÖFIT-Trendschau, einer stetig erweiterten Sammlung von mittlerweile 54 Trends und Themen. Identifiziert werden diese Trends und Themen durch interne und externe Expert:innen. In dieser Hinsicht stellt der quantitative Ansatz des Kapitels »Trendschau der Daten« der Publikation aus dem Jahr 2013 eine Besonderheit dar. Hierbei wurden anhand von mehr als 60.000 wissenschaftlichen Veröffentlichungen mithilfe einer Textanalyse zunächst Wortpaare ermittelt und mithilfe der Unterstützung durch Expert:innen schließlich die Innovationsfelder »Anything as a Service«, »Das Meer der Daten«, »Smart Grid« und »Drahtlose Sensornetzwerke« herausgearbeitet. Doch wie haben sich die Innovationsfelder seitdem entwickelt und was bringt die Zukunft? Um diese Fragen zu beantworten, haben FOKUS-Expert:innen die Innovationsfelder neu betrachtet.

3.1 DIE IDEE ANYTHING AS A SERVICE WIRD WEITERLEBEN

Die Idee, IT-Leistungen in Form von Diensten (Services) in der »Cloud« (über das Internet) bereitzustellen und zu nutzen, hat sich in den letzten Jahren stark weiterentwickelt und verbreitet. Dies entspricht dem analogen Trend der »Sharing Economy«, bei der ein zentraler Anbieter eine von Vielen genutzte Ressource, bspw. Carsharing, anbietet und managt. So wird bei Software as a Service (SaaS) beispielsweise keine Lizenz mehr verkauft, sondern gemietet und der entsprechende Dienst nutzungsabhängig abgerechnet (Pay-per-Use). In der Public Cloud werden Plattformen, Container und Infrastrukturen als Dienst (PaaS, CaaS, IaaS) bereitgestellt. Anything as a Service (XaaS) bezeichnet den Trend, dieses Prinzip auch auf andere IT-Leistungen zu übertragen, wie z. B. Business Process as a Service, Desktop as a Service, Security as a Service und viele andere mehr. Dienste können dabei gebündelt und gemeinsam bereitgestellt werden.

Seit 2013 sind viele Angebote verschiedener Dienste geschaffen worden. Sowohl die großen globalen Anbieter als auch kleine und mittlere regionale Anbieter stellen Dienste für unterschiedliche Sektoren bereit. Die grundlegenden Technologien, um Anything-as-a-Service-Angebote zu implementieren und zu betreiben, sind inzwischen ausgereift, (frei) verfügbar und vielfach auch standardisiert. Es ist aber auch klar zu beobachten, dass die Umstellung der Geschäftsmodelle von Kauf auf Miete für die Anbieter eine große Herausforderung ist, sowohl aus technischer, kaufmännischer als auch kultureller Sicht. Es muss nicht nur das bestehende System technisch migriert, sondern auch die Organisation selbst umgestaltet werden. Dies bremst eine größere Verbreitung dieses Trends. Generell ist eine Oligopolisierung zu beobachten. Es sind große globale Plattformanbieter entstanden, deren Markplätze auch die Dienste Dritter anbieten und gemeinsam vermarkten. Innerhalb der öffentlichen Verwaltung hingegen ist der Trend Anything as a Service bisher weniger sichtbar.

Zukünftig werden sich Angebote in der Public Cloud wohl schwerer etablieren, da die Aspekte Datenschutz und IT-Verfügbarkeit strategisch an Bedeutung gewinnen. Die Nutzer von Diensten über das Internet müssen klären, wie sie die Datenhoheit über die eigenen Daten sicherstellen und wie sie ihre eigene digitale Souveränität gewährleisten können, d. h., dass eigene Daten nicht abfließen oder Dritte den Geschäftsbetrieb beeinträchtigen können, bspw. auf Basis des CLOUD Acts. Entsprechend wird die Idee von Anything as a Service weiterleben, aber die Umsetzung differenzierter sein. Je nach Kritikalität der betroffenen Daten und Prozesse werden Dienste intern vor Ort aufgesetzt oder in der Cloud genutzt. Die Grundidee, Anything as a Service auf (internen) Plattformen anzubieten, Schnittstellen zu veröffentlichen (API-first Principle) und Dienste in Form von Selbstbedienung (Portal) durch die Nutzer selbstständig verwalten zu können, wird sich jedoch weiterentwickeln, um die kontinuierliche Automatisierung/Digitalisierung von Prozessen und Dienstleistungen zu ermöglichen, auch über Partnergrenzen hinweg.

3.2 VOM MEER ZUM OZEAN DER DATEN

Daten sind einerseits Produkt, andererseits Voraussetzung und Gegenstand digitaler Anwendungen. Das Volumen täglich verarbeiteter Daten steigt aufgrund der zunehmenden Verbreitung von digitalen Geräten und Diensten kontinuierlich. Das Meer der Daten hat sich in einen Ozean verwandelt. Zahlreiche neue Geschäftsmodelle bauen auf Daten auf und in diesem Zusam-

DATEN WERDEN ALS GRUNDLAGE FÜR DIE ZUKÜNFTIGE WETTBEWERBSFÄHIGKEIT VON UNTERNEHMEN UND STAATEN GESEHEN.

menhang werden Daten als Grundlage für die zukünftige Wettbewerbsfähigkeit von Unternehmen und Staaten gesehen. Große Potenziale liegen ebenfalls in den nicht-kommerziellen gemeinwohlorientierten Anwendungen.

In der ÖFIT-Trendschau von 2013 wurden mit Open Data und Big Data Analytics zwei wesentliche Trendfelder beschrieben. Die prognostizierten Chancen von Open Data haben sich in Deutschland nur teilweise erfüllt. Das liegt zum einen an Vorbehalten gegenüber dem freien und unkontrollierten Zugang zu Datenbeständen, sei es aus Angst vor Missbrauch (vorwiegend im öffentlichen Sektor) oder vor dem Verlust von Wettbewerbsvorteilen (im privaten Sektor). Zum anderen ist es auf die unterschiedlichen Geschwindigkeiten in der Digitalisierung zurückzuführen, wodurch Open-Data-Portale Datenbestände in der Breite oft nur lückenhaft abbilden können. In einigen Bereichen (z. B. für die kommerzielle App-Entwicklung) hat sich Open Data als fruchtbar erwiesen.

Big Data Analytics hat sich in vielen Anwendungsfeldern wie z. B. für Echtzeitdiagnosen im Katastrophenschutz oder für prädiktive Vorhersagen in Fertigung und Vertrieb etabliert. Dazu haben die Steigerung von Rechnerkapazitäten und die Weiterentwicklung technischer Verfahren der Extraktion, Integration und Analyse von Daten ebenso wie der Bedeutungsgewinn semantischer Textanalysen beigetragen. Die 2013 formulierte These »Müll rein – Müll raus« gilt dabei nach wie vor. Zwar erhöht ein großer Stichprobenumfang tendenziell die statistische Güte, für eine valide Interpretierbarkeit von Daten sind neben der Datenqualität aber auch die Kontextrückbindung und eine theoriegeleitete Auswertung notwendig.

Mit dem Trend »Meer der Daten« wurde auch auf die Relevanz von Dateninfrastrukturen und diensteorientierten Architekturen verwiesen. Der Aufbau von skalierbaren und robusten Datenarchitekturen und Datenplattformen ist nach wie vor hochaktuell. Im Hinblick auf Cloud Computing ist mit der Vorstellung von GAIA-X die Diskussion um den Aufbau eines souveränen digitalen Ökosystems, das sich an europäischen Datenschutzstandards orientiert, in den Mittelpunkt gerückt.

Eine konstante Herausforderung besteht darin, die Chancen der Datennutzung für das Gemeinwohl und die Risiken für Bürger:innen durch die Sammlung und (zweckfremde) Weiterverwertung personenbezogener Daten angemessen auszubalancieren. Insbesondere in smarten, mit Sensoren ausgestatteten Umgebungen wird die Abgrenzung von personenbezogenen und nichtpersonenbezogenen Daten zunehmend schwierig. Zukünftig wird es weiter darum gehen, Sicherheit, Datenschutz, Vertrauenswürdigkeit und die Qualität von immer komplexeren Datenstrukturen zu gewährleisten. Neu im Trend ist der Bedarf an großen Datenmengen für die Entwicklung von künstlicher Intelligenz. Organisationen werden folglich in Zukunft Datenkompetenz ausbauen müssen, was den Wettbewerb um Fachkräfte am Arbeitsmarkt weiter anheizen wird.

3.3 SMART GRID UND DEZENTRALE FLEXIBILITÄT

Zur besseren Überwachung und Steuerung von Stromerzeugung, -verbrauch, -speicherung und -verteilung sowie zur Unterstützung von Prozessen in der Energiewirtschaft ist eine informations- und kommunikationstechnische Vernetzung von Betriebsmitteln und Akteuren unerlässlich. So hat zwar der Anteil erneuerbarer Energien am Stromverbrauch mit knapp 38 Prozent in 2018 die Zielmarke von 35 Prozent für das Jahr 2020 bereits vorzeitig übertroffen, der zur Vermeidung von Netzengpässen aufgrund von Leistungsschwankungen erforderliche Netzausbau ist aber noch nicht erfolgt. Auch durch die Volatilität der erneuerbaren Energien wächst der Steuerungsbedarf beim Netzmanagement, um z. B. Netzengpässen entgegenzuwirken. Eine detaillierte Erfassung und Prognose der Netzsituation mittels Sensorik, IT-Infrastruktur und entsprechender

IN ZUKUNFT WERDEN WEITERE STRUKTURELLE VERÄNDERUNGEN DES ENERGIESYSTEMS ERWARTET.

Datenanalyse stellt nach wie vor eine Herausforderung dar. Insbesondere fehlt in Deutschland die flächendeckende Ausstattung des Netzes mit Smart-Meter-Systemen, deren Daten zeitlich und örtlich feinaufgelöste Informationen über Strombezug und -einspeisung liefern könnten

Durch Smart Meter sollen kurz- bis mittelfristig Informationen zum Nachfrageverhalten in Echtzeit bereitstehen und (ggf. durch Steuerungseingriff) dezentrale Flexibilität[105] erschlossen werden. Insgesamt wächst die Bedeutung eines smarten Energiemanagements weiter, das nicht nur auf den Abgleich von Energieverbrauch und -erzeugung fokussiert[106], sondern auch die beschränkte Kapazität der Stromnetze adressiert. Abregelungen sollen möglichst vermieden werden, stattdessen wird eine verstärkte Nutzung von Flexibilitäten, die sich z. B. durch den Ausbau der Elektromobilität ergeben, angestrebt. Die Integration erhöhter Einspeisung aus fluktuierenden dezentralen Quellen wie Wind und Photovoltaik erfordert allerdings eine verbesserte IT-Infrastruktur zur Steigerung der Flexibilität des Stromsystems als Ganzem.

Die Marktrolle der Verbraucher auf dem Strommarkt verändert sich mit der Digitalisierung. Für einen einzelnen Strom erzeugenden Haushalt (Prosumer)[107] gibt es deutliche Hürden am Strommarkt. So liegt sein Flexibilitätspotenzial oft unter dem geforderten Schwellenwert zur Teilnahme am Markt. Ist die Zahl wetterabhängiger, erneuerbarer dezentraler Energieversorger hoch, ist die Reaktionsfähigkeit der einzelnen Elemente des Energiesystems schwerer zu handhaben, da sie aufeinander abgestimmt werden müssen. Daher gibt es marktseitige Vorbehalte und technisch bedingte Grenzen. Die Digitalisierung und das Smart Metering unterstützen auf der technischen Ebene. Damit ist es nun zumindest möglich, dass sich einzelne dezentrale Stromversorger[108] zu sogenannten »Energie-Erzeuger-Gemeinschaften« auf lokaler Ebene zusammenschließen und gemeinsam als Akteur unmittelbar auf dem Strommarkt auftreten und wirtschaftlich am Energiemarkt teilnehmen. Speziell für solche Gemeinschaften werden digitale Komponenten konzipiert, die täglich auf Basis der verhaltensgenerierten, aggregierten, anonymisierten Smart-Meter-Daten der Haushalte Optimierungen und entsprechende Flexibilitäten ermitteln[109]. Während die technische Basis und das Interesse vorhanden sind, ist der Auftritt solcher Gemeinschaften auf dem Energiemarkt aus rechtlichen Gründen noch nicht in allen EU-Mitgliedsstaaten (z. B. auch nicht in Deutschland) für jeden Anwendungsfall möglich, was sich aber durch die neugefasste EU-Richtlinie zur Förderung von Erneuerbaren Energien (EE-Richtlinie)[110] ändern wird. Bürger:innen und auch Kommunen wird so das kollektive Zusammenwirken bei der Erzeugung und dem Verbrauch von Strom ermöglicht. Aufgrund der zunehmenden Digitalisierung der Energieinfrastrukturen und der Energiewirtschaft sowie dem Entstehen und der Erstarkung dezentraler Erzeuger-Gemeinschaften werden in Zukunft weitere strukturelle Veränderungen des Energiesystems erwartet.

[105] Dezentrale Flexibilität ist die Anpassungsfähigkeit von Erzeugungs-, Speicher- und Verbrauchsleistung dezentraler Anlagen im Stromnetz.

[106] Vergleiche S. 31 in Fromm, J.; Gauch, S.; Kaiser, T.; Weber, M. (2013).

[107] Siehe auch: Weber, M. (2019).

[108] Unter dezentraler Versorgung versteht man lokale, verbrauchsnahe Versorgungsformen mit elektrischer Energie, die die bestehende zentrale Versorgung ergänzen und ggf. ersetzen. Stromerzeugung und -speicherung erfolgen dabei durch eine Vielzahl kleiner, in eine Verteilinfrastruktur integrierter Geräte, die als verteilte Energieressourcen (DER: Distributed Energy Resources) bezeichnet werden.

[109] Vgl.: H2020 Projekt FLEXCoop – Democratizing energy markets through the introduction of innovative flexibility-based demand response tools and novel business and market models for energy cooperatives, siehe http://www.flexcoop.eu/.

[110] Richtlinie (EU) 2018/2001 des Europäischen Parlaments und des Rates vom 11. Dezember 2018, L 328/82, Amtsblatt der Europäischen Union vom 21. Dezember 2018. Online verfügbar unter: https://eur-lex.europa.eu/legal-content/DE/TXT/PDF/?uri=CELEX:32018L2001&from=DE.

DRAHTLOSE SENSORNETZWERKE HABEN SICH VON SPEZIALLÖSUNGEN ZU VERFÜGBAREN INFRASTRUKTUR-BAUSTEINEN WEITERENTWICKELT.

3.4 VON DRAHTLOSEN SENSORNETZWERKEN ZU IOT-FUNKNETZEN

Für eine automatisierte Unterstützung digitaler Prozesse sind vielfältige Informationen notwendig. Die physische Umgebung wird durch Sensoren erfasst, die als technische Sinnesorgane unterschiedlichste Parameter erfassen können. Die drahtlose Vernetzung verteilter Sensoren ist besonders komfortabel und spielt in verschiedenen »smarten« Szenarien – vom Smart Home bis zur Smart City – eine wichtige Rolle.

Seit der ÖFIT-Trendschau von 2013 sind Funknetze für das Internet der Dinge (Internet of Things, IoT)[111] als neue Infrastruktur verfügbar geworden, die sich insbesondere für die Anbindung von Sensoren eignet. Diese sogenannten Low Power Wide Area Networks (LPWAN)[112] können zwar nur kleine Datenmengen übertragen, sind dafür aber besonders energiesparsam und können hohe Reichweiten überbrücken bzw. gut Gebäude durchdringen. Preiswerter als klassischer Mobilfunk und einfacher nutzbar als spezielle Funklösungen steht damit eine neuartige technische Komponente für innovative Anwendungen und Produkte zur Verfügung. Typische Anwendungen sind Umweltsensoren, Tracker für Gegenstände oder Haustiere, Belegungssensoren für Parkplätze oder das Fernauslesen von Verbrauchszählern.

Das besprochene Innovationsfeld der drahtlosen Sensornetzwerke hat sich von Speziallösungen zu verfügbaren Infrastruktur-Bausteinen weiterentwickelt. Noch nicht so weit entwickelt ist die Vision von einer umfassenden Nutzung der Sensordaten über unterschiedliche Anwendungen hinweg, also die Datenquellen aus den Sensornetzen zu öffnen und zur Weiterverwendung in beliebigen Anwendungen bereitzustellen (bspw. in Form von Smart Data Hubs oder Open Data Portalen mit Echtzeitdaten).

Zukünftig wird es bei Sensor- bzw. IoT-Funknetzen weiter um die Verbesserung der Energieeffizienz bzw. die Energiegewinnung aus der Umwelt (Energy Harvesting) und die stärkere Integration gehen, sodass neue Einsatzgebiete erschlossen werden können und sich die Nutzung weiter vereinfacht. Die versprochenen Vorteile bei der Nutzung von Daten der Smart City werden sich auch daran messen lassen müssen, ob das Aufbrechen von Datensilos gelingt, sodass Daten aus öffentlichen und privaten Quellen zuverlässig und in ausreichender Qualität zur erweiterten Nutzung durch Dritte verfügbar werden.

[111] Mehr zum Internet der Dinge: Tiemann, J. (2019).

[112] Mehr zu LPWAN: Tiemann, J.; Manzke, F. (2019).

VORGEHEN

Vorbereitung der Datenanalyse

Welche Datenquellen eignen sich zur Ermittlung von Trends, die für öffentliche IT relevant sind? Anhand welcher Kriterien lassen sich solche Trends identifizieren? Die Untersuchung dieser Fragestellungen war der Ausgangspunkt der automatisierten Textanalyse. Für die Eignung einer Datenquelle sind etwa die Datenmenge, die Datenqualität, die Exportmöglichkeiten, die Aktualität und die Relevanz der verfügbaren Informationen für das Gebiet öffentliche IT von Bedeutung. Als Datenquellen wurden unter anderem Google Scholar[113], Scopus, IEEE Xplore[114] und das Web of Science[115] in Betracht gezogen, wobei letztlich trotz Schwächen bezüglich der Aktualität das Web of Science gewählt wurde. Hierbei waren insbesondere die Qualität der Daten und die vergleichsweise guten Exportmöglichkeiten ausschlaggebend.

Die Untersuchung der Eignung der Datenquellen war dabei nicht entkoppelt von der Suche nach Kriterien zur Trendidentifizierung, denn selbst das beste Kriterium ist nutzlos, wenn die verwendete Datenquelle keine passenden Informationen bereitstellt. Um die Fragestellung nach Kriterien zur Identifizierung von Trends zu bearbeiten, wurde eine Sammlung von Begriffen erstellt, die sich in vier Gruppen aufteilen ließ:

- Begriffe, die mit einem für öffentliche IT relevanten Trend verknüpft sind, beispielsweise »Social Bots«.
- Begriffe, die mit für öffentliche IT irrelevanten Trends verknüpft sind, etwa »CRISPR« aus dem Bereich der Biotechnologie.
- Begriffe, die keine Trends darstellen, beispielsweise der Kontinent »Australien«.
- Begriffe, die mit Themen und Technologien verknüpft sind, deren Bedeutung seit einiger Zeit stark nachlässt, etwa der Mobilfunkstandard »UMTS«.

Diese Sammlung ergab sich teilweise aus bereits bestehenden Sammlungen der ÖFIT-Trendforschung und teilweise aus interner Expertise, wobei auf die Vielfalt (etwa bezüglich verschiedener wissenschaftlicher Bereiche) der Begriffe geachtet wurde.

Für jeden Begriff dieser Sammlung wurden Daten zu wissenschaftlichen Publikationen aus den genannten Quellen exportiert und untersucht. Das Ziel waren quantitative Kriterien, anhand derer sich möglichst viele Trends möglichst frühzeitig identifizieren lassen, bei einer möglichst geringen Rate von falsch positiven Einordnungen. Auf Basis dieses Anspruchs wurden vier Eigenschaften[116] für die mit Trends der öffentlichen IT verbundenen Begriffen ermittelt:

- Die Begriffe treten bereits frühzeitig zumindest einmalig in den Titeln wissenschaftlicher Publikationen auf und nicht nur in den Abstracts.
- Die Begriffe treten besonders häufig in wissenschaftlichen Konferenzbeiträgen auf und vergleichsweise selten in Artikeln wissenschaftlicher Fachzeitschriften.
- Die Begriffe treten häufig in Publikationen auf, die den Bereichen Informatik oder Elektroingenieurwesen zugeordnet sind.
- Die Anzahl des Auftretens der Begriffe weist in den letzten zwei Jahren ein starkes Wachstum auf, während die Begriffe in den Jahren zuvor kaum erwähnt wurden.

Diese Eigenschaften wurden quantifiziert und flossen in Form von Grenzwerten in Algorithmen zur Trendidentifizierung ein. Aus diesen Grenzwerten wurde wiederum ein Bewertungssystem abgeleitet, das eine Rangordnung der Begriffe ermöglichte.

Datenanalyse

Als Betrachtungszeitraum wurde 2013 bis 2018 gewählt. Die Datengrundlage bestand aus fast zwei Millionen englischsprachiger Konferenzbeiträge zu mehreren tausend wissenschaftlichen Konferenzen pro Jahr. Die Daten wurden zunächst bereinigt. Beispielsweise wurden unvollständige Datensätze entfernt und für die Texte der Publikationen eine Stammformreduzierung durchgeführt. Anschließend wurden zunächst die Titel der Publikationen aus dem Jahr 2018 betrachtet und alle Wortfolgen aus ein bis vier Begriffen ermittelt. Die Menge an Wortfolgen wurde anhand der Kriterien und durch Hinzunahme der Publikationen aus früheren Jahren Stück für Stück verkleinert, bis schließlich 7012 Wortfolgen übrig waren, die alle Kriterien über den gesamten Betrachtungszeitraum erfüllten.

[113] https://scholar.google.com.

[114] https://ieeexplore.ieee.org.

[115] https://www.webofknowledge.com.

[116] Nicht alle, aber zumindest viele für den Bereich öffentliche IT relevante Trends weisen alle diese Eigenschaften auf.

1.947.973 PUBLIKATIONEN

29.291 KONFERENZEN

7.012 WORTFOLGEN

Innovationsfelder erkennen

Aus der Menge der 7012 Wortfolgen wurde anschließend eine Menge von etwas mehr als 200 Wortfolgen ausgewählt, die den höchsten Rang bezüglich des entwickelten Bewertungssystems aufwiesen. Für jede dieser Wortfolgen wurde dann eine Datenanalyse durchgeführt, deren visualisiertes Ergebnis zu einem Steckbrief zusammengetragen wurde. Diese Steckbriefe dienten Expert:innen als Hilfe zur Beurteilung, ob es sich bei einer Wortfolge tatsächlich um einen für die Öffentliche IT relevante Trend handelt und welche dieser Trends Anzeichen für ein bzw. Teil eines Innovationsfeldes sind. Als Ergebnis dieses Prozesses ergaben sich die fünf Innovationsfelder. Um das Ergebnis zu verifizieren und bisher nicht erfasste Entwicklungen zu berücksichtigen, wurde eine weitere Datenanalyse durchgeführt, nun allerdings auf Basis von Scopus statt des Web of Science und unter Berücksichtigung von Publikationen aus dem Jahr 2019. In der weiteren Bearbeitung der Innovationsfelder wurden zudem zusätzliche Datenquellen wie etwa soziale Medien und der Dokumentenserver arXiv berücksichtigt.

QUELLEN

Adi, Y.; Baum, C.; Cisse, M.; Pinkas, B.; Keshet, J. (2018): »Turning Your Weakness Into a Strength: Watermarking Deep Neural Networks by Backdooring«. Online verfügbar unter: https://arxiv.org/abs/1802.04633v3.

Agarwal, C.; Dong, B.; Schonfeld, D.; Hoogs, A. (2018): »An Explainable Adversarial Robustness Metric for Deep Learning Neural Networks«. Online verfügbar unter: https://arxiv.org/abs/1806.01477v2.

Akhtar, N.; Mian, A. (2018): »Threat of Adversarial Attacks on Deep Learning in Computer Vision: A Survey«. Online verfügbar unter: https://arxiv.org/abs/1801.00553v3.

Allen, R. (2017): »What happens online in 60 seconds?«. Online verfügbar unter: https://www.smartinsights.com/internet-marketing-statistics/happens-online-60-seconds/.

Armanious, K.; Jiang, C.; Fischer, M.; Küstner, T.; Nikolaou, K.; Gatidis, S.; Yang, B. (2019): »MedGAN: Medical Image Translation using GANs«. Online verfügbar unter: https://arxiv.org/abs/1806.06397v2.

Athalye, A.; Engstrom, L.; Ilyas, A.; Kwok, K. (2018): »Synthesizing Robust Adversarial Examples«. Online verfügbar unter: https://arxiv.org/abs/1707.07397v3.

Bachl, M.; Hartl, A.; Fabini, J.; Zseby, T. (2019): »Walling up Backdoors in Intrusion Detection Systems«. Online verfügbar unter: https://arxiv.org/abs/1909.07866v2.

Badjatiya, P.; Gupta, S.; Gupta, M.; Varma, V. (2017): »Deep Learning for Hate Speech Detection in Tweets«. Online verfügbar unter: https://arxiv.org/abs/1706.00188v1.

Bano, S.; Sonnino, A.; Al-Bassam, M.; Azouvi, S.; McCorry, P.; Meiklejohn, S.; Danezis, G. (2017): »Consensus in the Age of Blockchains«. Online verfügbar unter: https://arxiv.org/abs/1711.03936v2.

BBC (2019): »Fake voices ‚help cyber-crooks steal cash'«. Online verfügbar unter: https://www.bbc.com/news/technology-48908736.

Bessa, E. E.; Martins, J. S. B. (2019): »A Blockchain-based Educational Record Repository«. Online verfügbar unter: https://arxiv.org/abs/1904.00315v1.

Bündnis Bürgerenergie e. V. (2019): »Europäische Förderung von kollektiver Eigenversorgung und Erneuerbarer-Energie-Gemeinschaften«. Rechtliche Stellungnahme von Rechtsanwalt Dr. Philipp Boos für das Bündnis Bürgerenergie e. V. vom 22.08.2019. Online verfügbar unter: https://www.buendnis-buergerenergie.de/veroeffentlichungen/publikationen/.

Carlini, N.; Wagner, D. (2018): »Audio Adversarial Examples: Targeted Attacks on Speech-to-Text«. Online verfügbar unter: https://arxiv.org/abs/1801.01944v2.

Chang, Y.; Iakovou, E.; Shi, W. (2019): »Blockchain in Global Supply Chains and Cross Border Trade: A Critical Synthesis of the State-of-the-Art, Challenges and Opportunities«. Online verfügbar unter: https://arxiv.org/abs/1901.02715v1.

Chen, X.; Liu, C.; Li, B.; Lu, K.; Song, D. (2017): »Targeted Backdoor Attacks on Deep Learning Systems Using Data Poisoning«. Online verfügbar unter: https://arxiv.org/abs/1712.05526v1.

Dalí Museum (2019): »Dalí Lives: Museum Brings Artist Back to Life with AI«. Online verfügbar unter: https://thedali.org/press-room/dali-lives-museum-brings-artists-back-to-life-with-ai/.

Das, D.; Dutta, A. (2019): »Bitcoin's energy consumption: Is it the Achilles heel to miner's revenue?«. Online verfügbar unter: https://doi.org/10.1016/j.econlet.2019.108530.

Date, P.; Ganesan, A.; Oates, T. (2017): »Fashioning with Networks: Neural Style Transfer to Design Clothes«. Online verfügbar unter: https://arxiv.org/abs/1707.09899v1.

Davidson, T.; Warmsley, D.; Macy, M.; Weber, I. (2017): »Automated Hate Speech Detection and the Problem of Offensive Language«. Online verfügbar unter: https://www.aaai.org/ocs/index.php/ICWSM/ICWSM17/paper/download/15665/14843.

Djuric, N.; Zhou, j.; Morris, R.; Grbovic, M.; Radosavljevic, V.; Bhamidipati, N. (2015): »Hate Speech Detection with Comment Embeddings«. Online verfügbar unter: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.697.9571&rep=rep1&type=pdf.

Fang, M.; Yang, G.; Gong, N. Z.; Liu, J. (2018): »Poisoning Attacks to Graph-Based Recommender Systems«. Online verfügbar unter: https://arxiv.org/abs/1809.04127v1.

Fidel, G.; Bitton, R.; Shabtai, A. (2019): »When Explainability Meets Adversarial Learning: Detecting Adversarial Examples using SHAP Signatures«. Online verfügbar unter: https://arxiv.org/abs/1909.03418v1.

Finck, M. (2019): »Blockchain and the General Data Protection Regulation«. European Parliament, Science and Technology Options Assessment (STOA). Online verfügbar unter: https://europarl-eplibrary.hosted.exlibrisgroup.com/permalink/f/1iugl66/32EPA_EP_RESEARCH_DS835515/FULL/EN.

Frankle, J.; Carbin, M. (2019): »The Lottery Ticket Hypothesis: Finding Sparse, Trainable Neural Networks«. Konferenzpapier der ICLR 2019. Online verfügbar unter: https://openreview.net/pdf?id=rJl-b3RcF7.

Fromm, J.; Gauch, S.; Kaiser, T.; Weber, M. (2013): »ÖFIT-Trendschau: Innovationsfelder Öffentlicher IT«. Kompetenzzentrum Öffentliche IT. Online verfügbar unter: https://www.oeffentliche-it.de/publikationen?doc=14618&title=%C3%96FIT-Trendschau.

Gahntz, M. (2019): »Erklärbare KI«. In: Jens Fromm und Mike Weber, Hg., 2016: ÖFIT-Trendschau: Öffentliche Informationstechnologie in der digitalisierten Gesellschaft. Kompetenzzentrum Öffentliche IT. Online verfügbar unter: https://www.oeffentliche-it.de/-/erklarbare-ki.

Green, S. (2019): »Computer Code as Law: A New Frontier?«. JOTWELL (August 5, 2019). (Reviewing Karen Yeung, Regulation by Blockchain: the Emerging Battle for Supremacy between the Code of Law and Code as Law, 82 Mod. L. Rev. 207 (2019)). Online verfügbar unter: https://torts.jotwell.com/computer-code-as-law-a-new-frontier/.

Grosch, D. (2017): »Neuronale Netze«. In: Jens Fromm und Mike Weber, Hg., 2016: ÖFIT-Trendschau: Öffentliche Informationstechnologie in der digitalisierten Gesellschaft. Berlin: Kompetenzzentrum Öffentliche IT. https://www.oeffentliche-it.de/-/neuronale-netze.

Gu, T.; Dolan-Gavitt, B.; Garg, S. (2017): »BadNets: Identifying Vulnerabilities in the Machine Learning Model Supply Chain«. Online verfügbar unter: https://arxiv.org/abs/1708.06733v1.

Gu, T.; Liu, K.; Dolan-Gavitt, B.; Garg, S. (2019): »BadNets: Evaluating Backdooring Attacks on Deep Neural Networks«. Online verfügbar unter: https://ieeexplore.ieee.org/document/8685687.

Gumz, J. D.; Weber, M.; Welzel, C. (2019): »Anonymisierung: Schutzziele und Techniken«. Kompetenzzentrum Öffentliche IT. Online verfügbar unter: https://www.oeffentliche-it.de/publikationen?doc=100278&title=Anonymisierung+-+Schutzziele+und+Techniken.

Hakkani-Tür, D.; Tur, G.; Celikyilmaz, A.; Chen, Y.; Gao, J.; Deng, L.; Wang, Y. (2016): »Multi-Domain Joint Semantic Frame Parsing using Bi-directional RNN-LSTM«. Online verfügbar unter: https://pdfs.semanticscholar.org/d644/ae996755c803e067899bdd5ea52498d7091d.pdf.

Hasan, H. R.; Salah, K. (2019): »Combating Deepfake Videos Using Blockchain and Smart Contracts«. In IEEE Access (Volume: 7). Online verfügbar unter: https://ieeexplore.ieee.org/document/8668407.

Hassan, N. U.; Yuen, C.; Niyato, D. (2019): »Blockchain Technologies for Smart Energy Systems: Fundamentals, Challenges and Solutions«. Online verfügbar unter: https://arxiv.org/abs/1909.02914v1.

Hernandez-Farias, D. I.; Patti, V.; Rosso, P. (2016): »Irony Detection in Twitter: The Role of Affective Content«. ACM Transactions on Internet Technology. 16(3):19:1-19:24. doi:10.1145/2930663. Online verfügbar unter: https://riunet.upv.es/bitstream/handle/10251/81998/TOIT-autor.pdf?sequence=3.

Initiative Blockchain in der Verwaltung (2019): »Blockchain in der Verwaltung - Anwendungsbereiche und Herausforderungen«. Online verfügbar unter: http://bivd-initiative.de/wp-content/uploads/2019/08/Blockchain_in_der_Verwaltung_Teil_1_2019-08-30.pdf.

Jamriška, O.; Sochorová, Š.; Texler, O.; Lukáč, M.; Fišer, J.; Lu, J.; Shechtman, E.; Sýkora, D. (2019): »Stylizing Video by Example«. Online verfügbar unter: https://dcgi.fel.cvut.cz/home/sykorad/Jamriska19-SIG.pdf.

Kaji, S.; Kida, S. (2019): »Overview of image-to-image translation by use of deep neural networks: denoising, super-resolution, modality conversion, and reconstruction in medical imaging«. Online verfügbar: https://arxiv.org/abs/1905.08603v2.

Kakizaki, K.; Yoshida, K. (2019): »Adversarial Image Translation: Unrestricted Adversarial Examples in Face Recognition Systems«. Online verfügbar unter: https://arxiv.org/abs/1905.03421v2.

Kühl, E. (2018): »Auf Fake News folgt Fake Porn«. Die Zeit. Online verfügbar unter: https://www.zeit.de/digital/internet/2018-01/kuenstliche-intelligenz-deepfakes-porno-face-swap.

Li, J.; Li, N.; Peng, N.; Cui, H.; Wu, Z. (2019): »Energy consumption of cryptocurrency mining: A study of electricity consumption in mining cryptocurrencies«. Online verfügbar unter: https://doi.org/10.1016/j.energy.2018.11.046.

Ma, X.; Niu, Y.; Gu, L.; Wang, Y.; Zhao, Y.; Bailey, J.; Lu, F. (2019): »Understanding Adversarial Attacks on Deep Learning Based Medical Image Analysis Systems«. Online verfügbar unter: https://arxiv.org/abs/1907.10456v1.

Martinez-Hernandez, U.; Dehghani-Sanij, A. A. (2018): »Adaptive Bayesian inference system for recognition of walking activities and prediction of gait events using wearable sensors«. Online verfügbar unter: https://www.researchgate.net/publication/323534734_Adaptive_Bayesian_inference_system_for_recognition_of_walking_activities_and_prediction_of_gait_events_using_wearable_sensors.

Nakamoto, S. (2008): »Bitcoin: A peer-to-peer electronic cash system«. Online verfügbar unter: http://bitcoin.org/bitcoin.pdf.

Nguyen, C. T.; Hoang, D. T.; Nguyen, D. N.; Niyato, D.; Nguyen, H. T.; Dutkiewicz, E. (2019): »Proof-of-Stake Consensus Mechanisms for Future Blockchain Networks: Fundamentals, Applications and Opportunities«. In IEEE Access (Volume 7), S. 85727-85745. Online verfügbar unter: https://ieeexplore.ieee.org/document/8746079.

Nobata, C.; Tetreault, J.; Thomas, A.; Mehdad, Y.; Chang, Y. (2016): »Abusive Language Detection in Online User Content«. Online verfügbar unter: http://yichang-cs.com/yahoo/WWW16_Abusivedetection.pdf.

Olson, P. (2019): »Image-Recognition Technology May Not Be as Secure as We Think«. In The Wall Street Journal (Online-Version). Online verfügbar unter: https://www.wsj.com/articles/image-recognition-technology-may-not-be-as-secure-as-we-think-11559700300.

OpenAI (2019): »Better Language Models and Their Implications«. OpenAI Blog. Online verfügbar unter: https://openai.com/blog/better-language-models/.

Opiela, N.; Mohabbat Kar, R.; Thapa, B.; Weber, M. (2018): »Exekutive KI 2030«. Kompetenzzentrum Öffentliche IT. Online verfügbar unter: https://www.oeffentliche-it.de/documents/10181/14412/Exekutive+KI+2030+-+Vier+Zukunftsszenarien+f%C3%BCr+K%C3%BCnstliche+Intelligenz+in+der+%C3%B6ffentlichen+Verwaltung.

Paudice, A.; Muñoz-González, L.; Gyorgy, A.; Lupu, E. C. (2018): »Detection of Adversarial Training Examples in Poisoning Attacks through Anomaly Detection«. Online verfügbar unter: https://arxiv.org/abs/1802.03041v1.

Postinett, A. (2016): »Nadella ruft das Ende der Apps aus«. In Handelsblatt (Online-Version). Online verfügbar unter: https://www.handelsblatt.com/technik/it-internet/microsoft-setzt-auf-bots-nadella-ruft-das-ende-der-apps-aus/13383896-all.html.

Prabhumoye, S.; Tsvetkov, Y.; Salakhutdinov, R.; Black, A. W. (2018): »Style Transfer Through Back-Translation«. In »Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Long Papers)«, Seite 866 bis 876. Online verfügbar unter: https://www.aclweb.org/anthology/P18-1080.pdf.

Rangelov, D.; Tcholtchev, N.; Lämmel, P.; Schieferdecker, I. K. (2019): »Experiences Designing a Multi-Tier Architecture for a Decentralized Blockchain Application in the Energy Domain«. 11th International Congress on Ultra Modern Telecommunications and Control Systems and Workshops (ICUMT).

Reyes, A.; Rosso, P. (2013): »On the difficulty of automatically detecting irony: beyond a simple case of negation«. Online verfügbar unter: https://link.springer.com/article/10.1007/s10115-013-0652-8.

Saleh, F. (2019): »Blockchain Without Waste: Proof-of-Stake«. Online verfügbar unter: http://dx.doi.org/10.2139/ssrn.3183935.

Shafahi, A.; Huang, W. R.; Studer, C.; Feizi, S.; Goldstein, T. (2019): »Are adversarial examples inevitable?«. Online verfügbar unter: https://arxiv.org/abs/1809.02104v2.

Shao, G. (2019): »Fake videos could be the next big problem in the 2020 elections«. CNBC. Online verfügbar unter: https://www.cnbc.com/2019/10/15/deepfakes-could-be-problem-for-the-2020-election.html.

Shirasaki, M.; Yoshida, N.; Ikeda, S. (2019): »Denoising Weak Lensing Mass Maps with Deep Learning«. Online verfügbar unter: https://arxiv.org/abs/1812.05781v2.

Steinhardt, J.; Koh, P. W.; Liang, P. (2017): »Certified Defenses for Data Poisoning Attacks«. Online verfügbar unter: https://arxiv.org/abs/1706.03691v2.

Su, J.; Vargas, D. V.; Kouichi, S. (2019): »One pixel attack for fooling deep neural networks«. Online verfügbar unter: https://arxiv.org/abs/1710.08864v6.

Sultani, W.; Chen, C.; Shah, M. (2018): »Real-world Anomaly Detection in Surveillance Videos". Online verfügbar unter: http://openaccess.thecvf.com/content_cvpr_2018/papers/Sultani_Real-World_Anomaly_Detection_CVPR_2018_paper.pdf.

Tiemann, J. (2016): »Internet der Dinge«. In: Jens Fromm und Mike Weber, Hg., 2016: ÖFIT-Trendschau: Öffentliche Informationstechnologie in der digitalisierten Gesellschaft. Kompetenzzentrum Öffentliche IT. Online verfügbar unter: https://www.oeffentliche-it.de/-/internet-der-dinge.

Tiemann, J.; Manzke, F. (2019): »Funkende Dinge«. In: Jens Fromm und Mike Weber, Hg., 2016: ÖFIT-Trendschau: Öffentliche Informationstechnologie in der digitalisierten Gesellschaft. Berlin: Kompetenzzentrum Öffentliche IT. Online verfügbar unter: https://www.oeffentliche-it.de/-/funkende-dinge.

Tsipras, D.; Santurkar, S.; Engstrom, L.; Turner, A.; Madry, A. (2019): »Robustness May Be at Odds with Accuracy«. Online verfügbar unter: https://arxiv.org/abs/1805.12152v5.

Vincent, J. (2019): »Deepfake detection algorithms will never be enough«. The Verge. Online verfügbar unter: https://www.theverge.com/2019/6/27/18715235/deepfake-detection-ai-algorithms-accuracy-will-they-ever-work.

Wallace, B. C. (2013): »Computational irony: A survey and new perspectives«. Online verfügbar unter: https://link.springer.com/article/10.1007/s10462-012-9392-5.

Wang, W.; Wang, L.; Tang, B.; Wang, R.; Ye, A. (2019): »A Survey: Towards a Robust Deep Neural Network in Text Domain«. Online verfügbar unter: https://arxiv.org/abs/1902.07285v3.

Weber, M. (2017): »Prosument«. In: Jens Fromm und Mike Weber, Hg., 2016: ÖFIT-Trendschau: Öffentliche Informationstechnologie in der digitalisierten Gesellschaft. Kompetenzzentrum Öffentliche IT. Online verfügbar unter: https://www.oeffentliche-it.de/-/prosument.

Wei, S.; Saragih, J.; Simon, T.; Harley, A. W.; Lombardi, S.; Perdoch, M.; Hypes, A.; Wang, D.; Badino, H.; Sheikh, Y. (2019): »VR Facial Animation via Multiview Image Translation«. Facebook Research. Online verfügbar unter: https://research.fb.com/publications/vr-facial-animation-via-multiview-image-translation/.

Welzel, C.; Eckert, K.; Kirstein, F.; Jacumeit, V. (2017): »Mythos Blockchain: Herausforderung für den öffentlichen Sektor«. Kompetenzzentrum Öffentliche IT. Online verfügbar unter: https://www.oeffentliche-it.de/publikationen?doc=65740.

Yang, S.; Liu, J.; Lian, Z.; Guo, Z. (2016): »Awesome Typography: Statistics-Based Text Effects Transfer«. Online verfügbar unter: https://arxiv.org/abs/1611.09026v2.

Zakhary, V.; Agrawal, D.; El Abbadi, A. (2019): »Transactional Smart Contracts in Blockchain Systems«. Online verfügbar unter: https://arxiv.org/abs/1909.06494v1.

Zhang, S.; Zhang, X.; Chan, J.; Rosso, P. (2018): »Irony Detection via Sentiment-based Transfer Learning«. Online verfügbar unter: http://www.xiuzhenzhang.org/publications/ipm2018.zhang.irony.pdf.

Zhu, C.; Huang, W. R.; Shafahi, A.; Li, H.; Taylor, G.; Studer, C.; Goldstein, T. (2019): »Transferable Clean-Label Poisoning Attacks on Deep Neural Nets«. Online verfügbar unter: https://arxiv.org/abs/1905.05897v2.

Zhu, J.; Zhang, R.; Pathak, D.; Darrell, T.; Efros, A. A.; Wang, O.; Shechtman, E. (2018): »Toward Multimodal Image-to-Image Translation«. Online verfügbar unter: https://arxiv.org/abs/1711.11586v4.

KONTAKT

Jan Dennis Gumz
Kompetenzzentrum Öffentliche IT (ÖFIT)
Tel.: +49 30 3463-7173
Fax: +49 30 3463-99-7173
info@oeffentliche-it.de

Fraunhofer-Institut für
Offene Kommunikationssysteme FOKUS
Kaiserin-Augusta-Allee 31
10589 Berlin

www.fokus.fraunhofer.de
www.oeffentliche-it.de
Twitter: @OeffentlicheIT

ISBN: 978-3-9819921-7-5

kre | 19112 (Illustrationen: Martha Friedrich;
Fotos: pixabay.com, wikimedia.org)