Initiative D21

Kooperationspartner: KPMG

Unterarbeitsgruppe Algorithmen-Monitoring | Stand: 20. Juni 2019

DENKIMPULS DIGITALE ETHIK:

Transparenz und Nachvollziehbarkeit algorithmischer Systeme

AUTORINNEN Corinna Balkow (Initiative D21 e. V.), Dr. Irina Eckardt (Initiative D21 e. V. / KPMG)

MITWIRKENDE Aljoscha Burchardt (DFKI), Lena-Sophie Müller (Initiative D21 e. V.), Dr. Nora Schultz (Geschäftsstelle Deutscher Ethikrat), Prof. Barbara Schwarze (Initiative D21 e. V. / Kompetenzzentrum Technik-Diversity-Chancengleichheit e. V.), Mitarbeitende von KPMG AG und KPMG Law

- **Algorithmische Systeme spielen mittlerweile in allen Lebensbereichen eine relevante Rolle. Ziel muss es sein, dass diese Systeme von Menschen überprüfbar sind. Grundlage dafür ist Transparenz und Nachvollziehbarkeit.**
- **Transparenz beschreibt zunächst nur eine Informationsbereitstellung und führt noch nicht zwingend zu einem Verständnis. Mehr noch: zu viele Informationen können die Nachvollziehbarkeit einschränken. Dennoch ist Transparenz eine notwendige Grundlage für Nachvollziehbarkeit.**
- **Nachvollziehbar wird eine Information erst dann, wenn die Zusammenhänge, die zum Entstehen der Information geführt haben und ihre unmittelbaren Auswirkungen, von einem Menschen verstanden werden. Daraus folgt, dass Nachvollziehbarkeit einer zielgruppenspezifischen Informationsbereitstellung bedarf.**
- **Algorithmische Systeme müssen von Anfang an so gestaltet werden, dass sie grundsätzlich Nachvollziehbarkeit gewährleisten können. Je nach gesellschaftlicher Relevanz können Anforderungen an Transparenz und Nachvollziehbarkeit variieren.**

I. Einführung

Begriffsklärungen zu Transparenz und Nachvollziehbarkeit bei algorithmischen Systemen

Kaum eine Entwicklung hat unser Leben in den letzten Jahren so verändert, wie die Digitalisierung, also die Einführung von IT-Systemen in praktisch allen Lebensbereichen. Computerprogramme basieren auf unzähligen Algorithmen, die häufig verborgen im Hintergrund eine sehr breite Spanne von Aufgaben auf den unterschiedlichsten Ebenen erledigen. Es gibt Algorithmen zum Addieren von Ziffern, zum Anlegen einer Datei auf einer Festplatte, zum Sortieren von Listen oder zum Trainieren eines Maschinellen Lernsystems. Die Komplexität dieser algorithmischen Systeme zeigt sich unter anderem in der Länge des notwendigen Programmcodes, der von wenigen Zeilen bis hin zu ganzen Programmbibliotheken reichen kann. Im Folgenden bezieht sich der Begriff **„Algorithmische Systeme“** nicht nur auf den Programmcode, sondern auch auf die Prozesse der Datenauswahl und -bewertung, Entscheidungen zur Oberflächengestaltung und Zugangsmöglichkeiten für Menschen, die mit dem System interagieren.

Algorithmische Systeme werden längst in unzähligen Bereichen angewendet und erleichtern dabei eine Vielzahl der alltäglichen Aufgaben. Sie begleiten Menschen – oft unbemerkt - von der Aufnahme an einer bestimmten Schule oder Universität, über Bewerbungsverfahren, mögliche Beförderungen bis zur Beurteilung der Kreditwürdigkeit.[1] Dennoch herrscht oft Unwissen bei Menschen, die algorithmische Systeme bewusst oder unbewusst nutzen. Auf der einen Seite gibt es Misstrauen gegenüber algorithmischen Systemen[2], die bspw. Empfehlungen anzeigen, auf deren Basis Entscheidungen getroffen werden. Auf der anderen Seite herrscht oft blindes Vertrauen in die Technik.[3]

Algorithmische Systeme sind keine eigenständigen Subjekte, die von sich aus handeln.[4] Je häufiger sie als entscheidendes Element wahrgenommen werden und je weniger allgemein verständlich sie durch ihre Komplexität werden, desto stärker wächst Misstrauen und das Verlangen nach Kontrolle. Die wenigsten Menschen verstehen die Kriterien, nach denen die Systeme ihre Berechnungen durchführen. Um das Vertrauen in algorithmische Systeme zu stärken, wird daher gefordert, sie transparenter zu gestalten und dadurch nachvollziehbarer zu machen. Hier bieten algorithmische Systeme eine Chance, gerade solche Entscheidungen, die bisher nach „menschlichem Ermessen“ getroffen wurden, auf eine transparente, gerechte und verbindliche Basis zu stellen. Darin liegt auch ein großes Potential.

Der vorliegende Denkimpuls untersucht, in welchem Verhältnis **Transparenz** und **Nachvollziehbarkeit** für unterschiedliche Arten der Nutzung, sowie verschiedene Gruppen stehen. Dabei wurden relevante Fragestellungen identifiziert und sich mit diesen aus sozioökonomischer, technologischer und ethisch-rechtlicher Sicht auseinandergesetzt. Der Denkimpuls wurde im Rahmen der Unterarbeitsgruppe Algorithmen-Monitoring der Initiative D21 e. V. erarbeitet. Die vorgestellten Thesen für den Umgang mit Transparenz und Nachvollziehbarkeit tragen zu einer differenzierteren Debatte bei und initiieren eine breitere Diskussion zum Sinn und Zweck von Transparenz und Nachvollziehbarkeit. Ziel ist eine weitere Grundlage für einen nachhaltigen Umgang mit den Begriffen sowie den Maßnahmen Transparenz und Nachvollziehbarkeit zu schaffen.

Der Begriff der **Transparenz** beschreibt zunächst nur eine Informationsbereitstellung. Ergänzend enthält der Begriff der **Nachvollziehbarkeit** ein subjektives Element und hängt davon ab, ob Informationen in der Sprache sowie unter Berücksichtigung des Hintergrundwissens und der intellektuellen Fähigkeiten der Adressaten vorgetragen werden.[5] Insofern stellt Transparenz eine Voraussetzung für Nachvollziehbarkeit dar. Nachvollziehbar wird eine Information dann, wenn die Zusammenhänge, die zum Entstehen einer Information geführt haben und ihre unmittelbaren Auswirkungen von einem Menschen verstanden werden. Transparenz und Nachvollziehbarkeit können in einem Spannungsverhältnis zueinanderstehen: Die EU-Datenschutzgrundverordnung (DSGVO) verpflichtet Firmen auf Anfrage Auskünfte über eine Einzelperson zu geben. Dies ermöglicht Menschen, die über sie gespeicherten Daten anzufordern und einzusehen. Dabei müssen sie jedoch erwarten, eine umfangreiche Liste von Einzeldaten zu erhalten.[6] Eine solche Gestaltung mag zwar als Sicherung von Transparenz verstanden werden, gewährleistet aber keine Nachvollziehbarkeit dieser Daten für alle Betroffenen. Zu viele Informationen können zulasten der Nachvollziehbarkeit gehen. Dazu sind eine Kontextualisierung, eine Darstellung von Zusammenhängen und eine Erklärung von möglichen Auswirkungen in geeigneter Sprache und Umfang nötig. Dies erfordert eine zielgruppengerechte Aufbereitung der entsprechenden Informationen. Je nach Bereich der eigenen Expertise ist eine andere Art von Informationsbereitstellung erforderlich.

1 O’Neil, C. (2016): Weapons of Math Destruction; Crown Random House; weiterführend: https://weaponsofmathdestructionbook.com (letzter Abruf: 14.06.2019)

2 Vodafone (2016): Big Data; online verfügbar unter: https://www.vodafone-institut.de/wp-content/uploads/2016/01/VodafoneInstitute-Survey-BigData-en.pdf/ (letzter Abruf: 12.06.2019)

3 Fink,R.D.: Vertrauen in autonome Technik; online verfügbar unter: https://core.ac.uk/download/pdf/46915729.pdf (letzter Abruf: 12.06.2019)

4 OECD Principles on Artificial Intelligence; online verfügbar unter: https://www.oecd.org/going-digital/ai/principles/ (letzter Abruf: 11.06.2019)

5 Schmitt, A. (2005): Bedingungen gerechten Handelns. Motivations- und handlungstheoretische Grundlagen liberaler Theorien; Springer VS, S. 105

6 Nocun, Katharina (2018): Die Daten, die ich rief. Verlag: Bastei Lübbe; weiterführend: http://kattascha.de/worum-geht-es-im-buch-die-daten-die-ich-rief/ (Letzter Abruf: 14.06.2019)

Akteure innerhalb und außerhalb des algorithmischen Systems

Im Folgenden werden verschiedene Akteure innerhalb und außerhalb algorithmischer Systeme vorgestellt. Diese werden in der Darstellung auf vier Ebenen mit unterschiedlichen Zielstellungen in Bezug auf Transparenz und Nachvollziehbarkeit zusammengefasst: Menschen, die das algorithmische System wissentlich oder unwissentlich nutzen, werden als **Betroffene** bezeichnet. Unternehmen, die algorithmische Systeme entwickeln, testen und/oder vertreiben, gelten als **Gestaltende**. Hinzu kommen Personen, die über Aufträge und Einsätze von algorithmischen Systemen entscheiden (**Entscheidende**) sowie Expertinnen und Experten unterschiedlicher Fachrichtungen als (externe) **Prüfende** eines algorithmischen Systems.

Betroffene Personen

Personen können als aktive Nutzende eines algorithmischen Systems oder als indirekte Objekte der Datenerzeugung betroffen sein. Aktive Nutzung würde dann beispielsweise die Nutzung einer Software für die Routenplanung betreffen. Indirekt betroffen sind Personen, wenn beispielsweise Bewegungsdaten aus Mobilgeräten für die Routenplanung in der entsprechenden Software genutzt werden.

In Bezug auf **betroffene Personen** wird Nachvollziehbarkeit an der Verständlichkeit gemessen. Dabei werden Maßnahmen zur Sicherstellung von Verständlichkeit im Allgemeinen für eine Mehrheit der Menschen in der jeweiligen Zielgruppe geplant. Besonders zu berücksichtigen sind darüber hinaus die Teilhabemöglichkeiten von Menschen mit besonderen Anforderungen an Erklärungen, beispielsweise von Kindern und Menschen mit Behinderungen.[7] Maßgefertigte Systeme für bestimmte Zielgruppen, z. B. Fachsoftware für medizinisches Personal, brauchen zunächst nur für diese nachvollziehbar sein. Eine Erklärung in medizinischer Fachsprache ist beispielsweise vorrangig für medizinisches Fachpersonal hilfreich. Zum weiterführenden Verständnis von Informationen über den Einsatz algorithmischer Systeme ist es förderlich, allgemein zugängliche Weiterbildungen[8] anzubieten, und Schulbildung um mehr statistische und informatische Grundlagen[9] zu erweitern.

Im direkten Kontakt ist ein erster Schritt zur Nachvollziehbarkeit eine transparente Darstellung des Einsatzes von algorithmischen Systemen angereichert um eine kurze Erläuterung. Dies könnte beispielsweise analog zur gängigen Formulierung „Dieses Schreiben wurde automatisch generiert und ist ohne Unterschrift gültig." erfolgen. So sollte beispielsweise in Chats klar kenntlich gemacht werden, welche Fragen von Chatbots beantwortet werden und wie eine Person erreicht werden kann. In bestimmten Situationen müssen Informationen schnell erfasst werden können, beispielsweise während eines (Online-)Einkaufs. Eine Vereinfachung der Darstellung und ein hoher Grad an Vergleichbarkeit dieser Informationen kann durch Strukturierung ermöglicht werden.[10]

Nach dem Einsatz eines algorithmischen Systems könnten Betroffenen die Kriterien für die Bewertung, ihre Gewichtung und die Daten, die der Berechnung des Ergebnisses zugrunde liegen, offengelegt werden, wenn es sich um teilhaberelevante Entscheidungen über Menschen handelt. Wenn ein algorithmisches System beispielsweise die Kreditwürdigkeit einer Person berechnet, sollte diese Person verstehen können, welche Daten sich wie auf das Ergebnis auswirken.

Darüber hinaus könnte eine externe Stelle zur Prüfung bei Verdacht auf Diskriminierung von Betroffenen angefragt werden, da das für eine solche Beurteilung notwendige Wissen nicht allgemein vorausgesetzt werden kann. Es muss in teilhaberelevanten Bereichen über ein systeminternes Beschwerdemanagement hinaus, eine konkrete externe Einspruchsmöglichkeit geben, um sicherzustellen, dass Widerspruchsmöglichkeiten für alle Menschen erreichbar sind.

7 Bundeszentrale für politische Bildung (2018): Allgemeine Geschäfts-Bedingungen der bpb in leicht verständlicher Sprache; online verfügbar unter: https://www.bpb.de/shop/201038/allgemeine-geschaefts-bedingungen in leichter Sprache/ (letzter Abruf: 08.05.2019)
8 Elements of AI: Free online course; online verfügbar unter: https://www.elementsofai.com/ (letzter Abruf: 08.05.2019)
9 GI (2019): GI begrüßt Forderung nach Informatik-Pflichtfach in Niedersachsen; online verfügbar unter: https://gi.de/meldung/gi-begruesst-forderung-nach-informatik-pflichtfach-in-niedersachsen/ (letzter Abruf: 08.05.2019)
10 Transparency International Deutschland e.V.: Initiative Transparente Zivilgesellschaft; online verfügbar unter: https://www.transparency.de/mitmachen/initiative-transparente-zivilgesellschaft/ (letzter Abruf: 08.05.2019)

Gestaltende Personen

Auf der Ebene der **Gestaltenden** stehen Personen oder Unternehmen, die ein algorithmisches System konzipieren, entwickeln, testen und/oder auch vertreiben. Unternehmen in diesem Bereich werden bei Anforderungen an Transparenz und Nachvollziehbarkeit zwischen Aufwand und Nutzen abwägen. Kosten für aufwändigere Verfahren müssen gerechtfertigt werden. Eigene Entwicklungen werden als schützenswerte Geschäftsgeheimnisse angesehen. Eine interne Qualitätskontrolle kann auf Transparenz im eigenen geschäftlichen Interesse bestehen. Daten, Algorithmen und andere Faktoren müssen für interne Tests verfügbar sein. Zwischen den einzelnen Phasen in der Gestaltung eines algorithmischen Systems besteht ein Interesse an nachvollziehbaren Informationen. Eine solche Nachvollziehbarkeit innerhalb des Systems dient der Kontrolle der Funktionalität und kann auch zur Verbesserung genutzt werden.

Entscheiderinnen und Entscheider

Unter **entscheidenden Personen** werden Personen oder Organisationen zusammengefasst, die als rechtlich, technisch, oder politisch Verantwortliche entscheiden, welche algorithmischen Systeme zu welchem Zweck eingesetzt und wie geprüft werden. Um ausgewogene Entscheidungen zu treffen und Leitlinien, Regulierungen, Gesetze für Gestaltende algorithmischer Systeme zu entwickeln, benötigen sie interdisziplinären Austausch, sowie die Einbindung von unterschiedlichen kulturellen Hintergründen.[11] Beispiele für Regulierungen betreffen beispielsweise, wie datensparsam oder datenintensiv algorithmische Systeme geplant werden. Eine spätere Änderung ist dahingegen deutlich schwieriger und teurer. Analog zu „Privacy by Design“ können Anforderungen an „Transparency by Design“[12] formuliert werden.

Die Anforderung auf der Ebene der Entscheidenden besteht in einer Nachvollziehbarkeit der Prozesse, Ergebnisse und Beschränkungen algorithmischer Systeme. Eine Pflicht zur Durchführung externer Prüfungen würde zu mehr Nachvollziehbarkeit führen. Für die Gesellschaft bleibt eine „Balance zwischen Offenheit und Geheimhaltung wichtig“.[13] Es müssen also die Rechte der verschiedenen Akteure identifiziert und gegebenenfalls geschützt werden. Betroffene können beispielweise in bestimmten Situationen ein Recht auf Geheimhaltung haben.[14]

Extern prüfende Personen

Aus der Sicht **externer Prüfender** müssen vielfältige Faktoren beachtet werden. Extern bedeutet hier, dass die Personen nicht Teil der Personengruppe sind, die das algorithmische System gestaltet. Sondern sie prüfen das System als externe Dritte von außen. Für eine umfassende Prüfung sollten relevante Daten, Algorithmen, Modelle und Prozesse offengelegt werden, damit mögliche Fehlentscheidungen und Bias/Diskriminierungen aufgedeckt werden können. Eine Nachvollziehbarkeit algorithmischer Systeme ist nur in interdisziplinärer Arbeit erzielbar: Beispielsweise würden technische Expertinnen und Experten dazu den Quellcode betrachten, rechtliche das nachvollziehbare Verhalten beurteilen und sprachliche die Erläuterung leicht verständlich formulieren. Die Ergebnisse externer Prüfungen sollten veröffentlicht werden.[15] Externe Prüfende umfassen auch diejenigen, die prüfen, inwieweit Gesetze und Regulierungen umgesetzt werden.

11 Balkow, C., Eckardt, I (2019): Bias in Algorithmischen Systemen: online verfügbar unter: https://initiatived21.de/app/uploads/2019/03/algomon_denkimpuls_bias_190318.pdf (letzter Abruf 11.06.2019)

12 Mascharka, et.al (2018): Transparency by Design: Closing the Gap Between Performance and Interpretability in Visual Reasoning; online verfügbar unter: http://openaccess.thecvf.com/content_cvpr_2018/papers/Mascharka_Transparency_by_Design_CVPR_2018_paper.pdf (letzter Abruf: 23.05.2019)

13 Denk, Felix (2016): Wann Algorithmen transparent sein sollten – und wann nicht; online verfügbar unter: https://www.fluter.de/kuenstliche-intelligenz-ethik-algorithmen-kontrollieren/ (letzter Abruf: 08.05.2019)

14 Duttge, G., et.al. (2015): Normatives Fundament und anwendungs-praktische Geltungskraft des sogenannten Rechts auf Nichtwissen; online verfügbar unter: http://www.recht-auf-nichtwissen.uni-goettingen.de/ (letzter Abruf: 13.06.2019)

15 Semsrott, A. (2019): Verwaltungsgerichte entscheiden für Transparenz; online verfügbar unter: https://fragdenstaat.de/blog/2019/05/21/topf-secret-klagewelle/ (letzter Abruf: 13.06.2019)

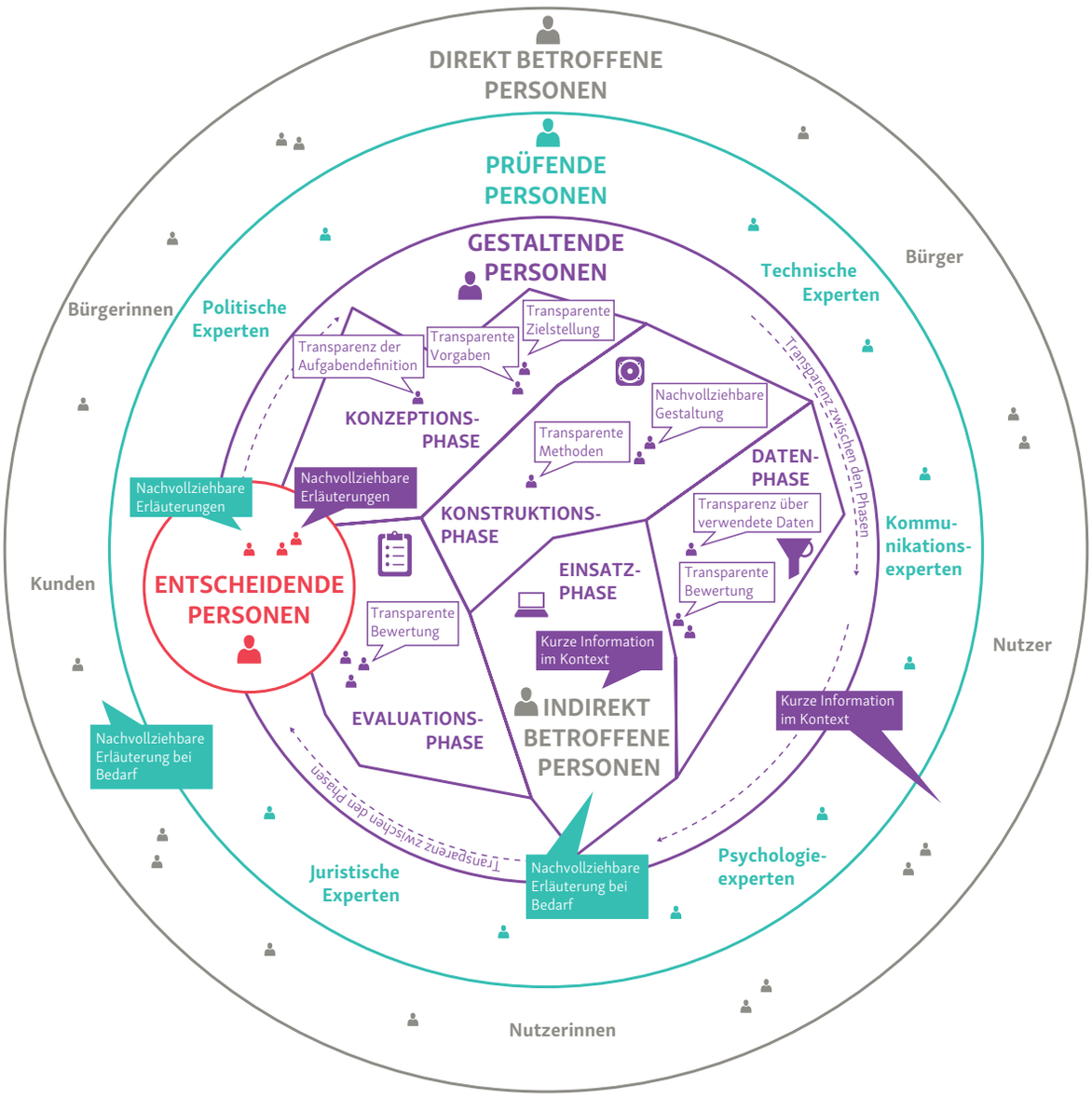

Abbildung 1: Zusammenhang von Transparenz und Nachvollziehbarkeit in Bezug auf die vier genannten Gruppen innerhalb und außerhalb eines beispielhaften algorithmischen Systems.

Gestaltende algorithmischer Systeme benötigen Transparenz innerhalb des Systems

Im inneren Kreis der Grafik wird der Entwicklungszyklus eines algorithmischen Systems dargestellt. In der Konzeptionsphase werden Anforderungen an Ausgaben und Zielsetzungen definiert. In der Konstruktionsphase werden diese Vorgaben umgesetzt. In der Datenphase werden Daten gesammelt, bewertet und für den Einsatz aufbereitet. In der Einsatzphase kommen Betroffene Personen mit dem System direkt in Berührung. In der Evaluationsphase werden die Ergebnisse bewertet. In allen Entwicklungsphasen befinden sich **Gestaltende** eines algorithmischen Systems. Zwischen den Phasen kann ein transparenter Austausch von Informationen stattfinden. Oft reichen jedoch für Absprachen auch nachvollziehbare Erläuterungen.

Entscheidende müssen Auswirkungen und Konsequenzen algorithmischer Systeme nachvollziehen können

Um die verschiedenen Implikationen algorithmischer Systeme einschätzen zu können, müssen **Entscheidende** sowohl untereinander als auch mit den anderen Akteuren

kommunizieren. Je nach eigener Expertise benötigen sie transparente Informationen oder eine nachvollziehbare Aufbereitung durch Fachleute anderer Disziplinen bzw. Spezialisierungen. Je nach Einsatzkontext muss entschieden werden, ob selbstlernende Algorithmen eingesetzt werden, deren Ausgaben schwerer nachzuvollziehen sind. Bei der Einschätzung algorithmischer Systeme sind nicht nur die Algorithmen zu berücksichtigen, sondern auch die Datenquellen, die Implementation und die Art des späteren Einsatzes. Es muss kommuniziert werden, inwieweit das algorithmische System am Einsatzort als Unterstützung (und nicht als alleinige Grundlage für Entscheidungen) verwendet werden kann.

Externe Prüfende benötigen Einsicht in den Entwicklungszyklus des algorithmischen Systems

Externe Prüfende stehen teilweise zwischen Gestaltenden, Entscheidenden und Betroffenen. Für eine aussagekräftige Prüfung benötigen sie Transparenz über den gesamten Lebenszyklus eines algorithmischen Systems, die Prozesse der Entscheidungsfindung und die Bedürfnisse der Betroffenen. Sie übermitteln eine nachvollziehbare Einschätzung an die Auftraggeber oder Auftraggeberin der Prüfung – Entscheidende oder Betroffene. Dabei wird das algorithmische System z. B. auf mögliche Diskriminierungen untersucht. Durch eine Veröffentlichung von Prüfberichten kann Nachvollziehbarkeit ohne vollständige Transparenz des algorithmischen Systems erreicht werden.

Betroffene benötigen nachvollziehbare Aufbereitung und Erläuterung von Informationen

Direkt Betroffene befinden sich als aktive Nutzende rund um das algorithmische System. Damit direkt Betroffene einschätzen können, wann sie mit algorithmischen Systemen in Berührung kommen, brauchen sie eine einfache Kennzeichnung im Kontext der Anwendung. Dies könnte durch die Nutzung von Symbolen ermöglicht werden. Bei Bedarf – zum Beispiel um zu verstehen, ob sie zu Unrecht von algorithmischen Systemen beeinträchtigt werden - muss ihnen eine nachvollziehbare Erläuterung zum Entscheidungsprozess ermöglicht werden.

Durch die allgemeine Verbreitung algorithmischer Systeme, sind indirekt Betroffene auch ohne aktive Teilnahme betroffen von Entscheidungen zum Einsatz solcher Systeme z.B. automatisierte Grenzkontrollen.[16] Indirekt Betroffene sind unwissentlich Teil des Systems, weil ihre analog erhobenen Daten digitalisiert wurden oder digitale Zugangsmöglichkeiten fehlen. Es gilt zu überlegen, wie indirekt Betroffene informiert werden können, wenn algorithmische Systeme zum Einsatz kommen. So könnte beispielsweise dargestellt werden, wann und wie personenbezogene Daten indirekt Betroffener als Trainingsdaten von Gestaltenden für die Verbesserung des algorithmischen Systems genutzt werden.

Unterschiedliche Anforderungen an Transparenz und Nachvollziehbarkeit

Im Folgenden werden die relevanten Fragestellungen bezüglich der unterschiedlichen Akteure aus drei Perspektiven betrachtet: technologisch, sozioökonomisch und ethisch-rechtlich. Dabei bezieht sich die technologische Perspektive auf die praktische Umsetzbarkeit von Anforderungen an Transparenz und Nachvollziehbarkeit und setzt sich mit den Bedingungen, Problemen und Möglichkeiten auseinander. Die sozioökonomische Perspektive arbeitet heraus, welche sozialen und ökonomischen Chancen und Herausforderungen durch die Anforderungen an Transparenz und Nachvollziehbarkeit algorithmischer Systeme entstehen und wie Herausforderungen begegnet werden kann. Die ethisch-rechtliche Perspektive behandelt die Erschließung rechtlicher Grundlagen, und eine mögliche Regulierung algorithmischer Systeme anhand ihres Diskriminierungs- und Schadenspotential.

16 Automatisiertes Grenzkontrollsystem EasyPASS; online verfügbar unter: https://www.easypass.de (Letzter Abruf 11.06.2019)

II. Technologische Perspektive auf Transparenz und Nachvollziehbarkeit in algorithmischen Systemen

Die Forderungen für ein besseres Verständnis von algorithmischen Systemen nehmen aufgrund weitreichender Verbreitung stetig zu. Allgemein vorherrschendes Misstrauen gegenüber umfangreicher Datensammlung[17], unergründbare Ergebnisse von algorithmischen Systemen[18] und durch künstliche Intelligenz unterstützte Entscheidungen[19] führen oft zur Forderung nach verpflichtender Transparenz. Transparenz führt wie oben dargelegt nicht automatisch zu Nachvollziehbarkeit. Mehr Transparenz von algorithmischen Systemen kann aber den Zugang zu relevanten Informationen erleichtern, die Nachvollziehbarkeit der Ergebnisse daher prinzipiell ermöglichen und könnte prinzipiell eine bessere Evaluation der Systeme erlauben. Eine zielgruppengerechte Aufbereitung dieser Informationen würde de facto zu mehr Nachvollziehbarkeit führen und so eine bessere Evaluation der Systeme erlauben. Transparenz ist also nötig für Kontrolle und Nachvollziehbarkeit dient als Voraussetzung für Vertrauen.

These: Transparenz kann der Verbesserung von algorithmischen Systemen dienen.

Beschreibung: Aus technologischer Sicht ist Transparenz - mit gewissen Einschränkungen im maschinellen Lernen - technisch herstellbar. Fraglich ist jedoch, ob diese Transparenz in Nachvollziehbarkeit umgewandelt werden kann. Als Zwischenstufe bedeutet Erklärbarkeit, dass ein Experte erkennen kann, was in einem algorithmischen System passiert. „Explainable AI" (abgekürzt XAI[20]) beschreibt die Forschungsrichtung, deren Ziel es ist, Entscheidungen einer künstlichen Intelligenz – eines algorithmischen Systems - für den Menschen einfach nachvollziehbar zu machen, d. h. zu verdeutlichen, warum und wie in algorithmischen Systemen bestimmte Faktoren gewichtet wurden und welche Annahmen der Entscheidung zugrunde liegen. Das algorithmische System muss also von Anfang an so geplant werden, dass Daten für die Analyse ausgegeben werden. Vorbedingung ist eine dem System immanente Implementierung von entsprechenden Ausgaben.

Beispiel: Forschungen zu erklärbarer KI betonen den Nutzen bei der Beurteilung, wie erfolgreich ein algorithmisches System gelernt hat. Beispielsweise zeigten Forscher der TU Berlin, dass KI-Systeme oft anhand von nicht für Menschen einsichtigen Merkmalen lernen. So wurden beispielsweise Boote und Züge nicht anhand ihrer eigenen Stuktur erkannt, sondern anhand des Hintergrunds: Boote wurden erkannt, wenn auf den Fotos Wasser war. Züge wurden daran erkannt, dass Schienen auf dem Foto waren. Bilder von Pferden stammten aus einer bestimmten Datenbank und wurden durch das algorithmische System anhand eines Copyright Merkmals erkannt.[21] Beim Thema möglicher Diskriminierung stellt sich die technische Frage, ob im System bestimmte Merkmale gekennzeichnet werden, die als Ersatz für verbotene persönliche Merkmale dienen, z. B. wenn bei Bewerbungen von Frauen anhand von Lücken im Lebenslauf auf Elternschaft geschlossen wird.

Umgang: Bisher eingesetzte algorithmische Systeme verfügen kaum über Ausgaben über ihre Berechnungswege. Stattdessen werden sie im Hintergrund eingesetzt.[22] Die Datenschutzgrundverordnung fordert jedoch erklärende Informationen.[23] Ein Beschwerdemanagement kann zu Hinweisen führen, wo Verbesserungen sinnvoll sind. Bei

17 Joler, V., Petrovski, A. (2016): Immaterial Labour and Data harvesting; online verfügbar unter: https://labs.rs/en/facebook-algorithmic-factory-immaterial-labour-and-data-harvesting/ (letzter Abruf 12.06.2019)
18 Prof. Dr. Zweig, K. et. al (2018): Wo Maschinen irren können; online verfügbar unter https://www.bertelsmann-stiftung.de/fileadmin/files/BSt/Publikationen/GrauePublikationen/WoMaschinenIrrenKoennen.pdf (letzer Abruf 12.06.2019)
19 Canon, G. (2019): How Taylor Swift showed us the scary future of facial recognition; online verfügbar unter: https://www.theguardian.com/technology/2019/feb/15/how-taylor-swift-showed-us-the-scary-future-of-facial-recognition/ (letzter Abruf: 08.05.2019)
20 Eisenstadt, V., Althoffl, K.-D., (2018): A Preliminary Survey of Explanation Facilities of AI-Based Design Support Approaches and Tools, online verfügbar unter: https://www.dfki.de/fileadmin/user_upload/import/9983_LWDA_2018_paper_59.pdf (letzter Abruf 12.06.2019)
21 Lapuschkin, Sebastian et al. (2019): Unmasking Clever Hans predictors and assessing what machines really learn; online verfügbar unter: https://www.nature.com/articles/s41467-019-08987-4.pdf/ (letzter Abruf: 08.05.2019)
22 West, S.M., Whittaker, M. and Crawford, K. (2019). Discriminating Systems: Gender, Race and Power in AI. AI Now Institute. Online verfügbar unter: https://ainowinstitute.org/ discriminatingsystems.html (letzter Abruf: 14.06.2019)
23 Selbst, Andrew D; Powles, Julia (2017): Meaningful information and the right to explanation; online verfügbar unter: https://doi.org/10.1093/idpl/ipx022/ (letzter Abruf: 08.05.2019)

Verschlüsselungstechniken gilt die Offenlegung des Quelltextes als Qualitätsmerkmal.[24]

These: Eine Veröffentlichung zu vieler Daten kann einer Nachvollziehbarkeit im Wege stehen.

Beschreibung: Sowohl durch eine stetig steigende Nutzung technologischer Lösungen als auch durch eine stetig steigende Speicherkapazität, stehen heutzutage mehr Daten zur Verfügung als je zuvor. Die darauf aufbauenden algorithmischen Systeme können immer komplexere Aufgaben bewältigen, benötigen hierfür jedoch auch eine große Menge an relevante Daten. Die Aufbereitung von Daten, bei der diese bereinigt, geordnet oder ergänzt werden, umfasst ca. 80 Prozent des Aufwandes beim Training neuronaler Netze. Diese wichtige Aufgabe wird oft ausgelagert[25], was gegebenenfalls zusätzliche Herausforderungen für eine nachvollziehbare Analyse mit sich bringt. Bei der geforderten Nachvollziehbarkeit stehen Genauigkeit und Verständlichkeit in einem Spannungsfeld. Je mehr technische Details man offenlegt, vielleicht bis auf die Grundlagen der Hardware oder Software, desto schwieriger wird die Nachvollziehbarkeit für die meisten Betroffenen. Beispiel: Wer in einem Online Shop einkauft, muss diverse persönliche Informationen angeben: Name, Adresse, Geburtsdatum, Kontoverbindung. In den Datenschutzbestimmungen ist ausführlich ausgeführt, was für welchen Zweck verwendet wird. Wenn nun die Daten der Anbieter abgefragt werden, erhalten die Anfragenden zum Teil ausführliche Excel-Dokumente inklusive Systemdaten, Gerätetypen und genutzten Betriebssystemen.[26]

Umgang: Je nach Zielgruppe und Kontext soll eine Abstufung in verschiedene Level von Transparenz erfolgen: Betroffene Personen: Menschen, die das algorithmische System im Alltag verwenden, benötigen meist weniger Detailtiefe. Eine wichtige Information, die auch schnell erfasst werden kann, wäre eine klare Kennzeichnung, dass ein algorithmisches System eingesetzt wurde, kombiniert mit einer kurzen Erläuterung. Dies kann ergänzt werden um Kerninformationen über seine Art oder Funktion.
27 Einfache Erläuterungen zum angezeigten Ergebnis erleichtern das Verständnis.[28] Bei Bedarf muss eine ausführliche Erläuterung einfach auffindbar sein. Schließlich sollte ein Einspruch gegenüber fehlerhaften Ergebnissen möglich sein.[29] Als Lösung kommen also sowohl Optionen in Frage, die eine möglichst große Informationsvielfalt bieten, die einer interessierten Nutzerin alle möglichen Fragen beantworten würden, als auch Kurzfassungen, die ähnlich wie eine Lebensmittelkennzeichnung einen kurzen Überblick erlauben.[30]

Entscheidende Personen: Menschen, die algorithmische Systeme beauftragen oder einsetzen, benötigen nachvollziehbare Erläuterungen für alle Schritte im algorithmischen System. Durch eine Zertifizierung von Prozessen in der Planung, Entwicklung und im Einsatz können auch nicht vollständig transparente algorithmische Systeme bewertet werden.[31]

Prüfende Personen: Menschen, die algorithmische Systeme entwickeln, testen oder prüfen, benötigen möglichst viele Informationen, um abzuschätzen, wie sicher das System zu Ergebnissen kommt.

24 Pillitteri, V.; Lightman, S. (2015) Guide to Industrial Control Systems (ICS) Security; online verfügbar unter: https://nvlpubs.nist.gov/nistpubs/specialpublications/nist.sp.800-82r2.pdf (letzter Abruf: 14.06.2019)
25 Stouffer, Keith et. al (2015): The invisible workers of the AI era - A new type of blue-collar industry has emerged around curating the data that powers AI; online verfügbar unter: https://towardsdatascience.com/the-invisible-workers-of-the-ai-era-c83735481ba/ (letzter Abruf: 08.05.2019)
26 Heidrich, J; Maekeler, N. (2019): Antwortet uns! DSGVO-Datenauskunft im Selbsttest; c´t 2019, Heft 13, S.171
27 Bier, C. (2018) Umsetzung des datenschutzrechtlichen Auskunftsanspruchs auf Grundlage von Usage-Control und Data-Provenance-Technologien; online verfügbar unter: http://publica.fraunhofer.de/dokumente/N-484529.html (letzter Abruf: 14.06.2019)
28 Ortloff, A.M. et al (2018): Evaluation kontextueller Datenschutzerklärungen; online verfügbar unter: http://publica.fraunhofer.de/dokumente/N-528032.html (letzter Abruf: 14.06.2019)
29 Warscheid, L. (2019): Ein Blitzer wird zum Fall für die Richter; online verfügbar unter: https://www.saarbruecker-zeitung.de/saarland/verfassungsgericht-verhandelt-ueber-blitzer-messung-aus-friedrichsthal_aid-38686685
30 Standardisierte Lizenzverträge; online verfügbar unter: https://de.creativecommons.org/ (letzter Abruf: 08.05.2019)
31 KI Verband (2019): KI-Gütesiegel – AI Made in Germany; online verfügbar unter: https://ki-verband.de/ki-guetesiegel-ai-made-in-germany/ (letzter Abruf: 08.05.2019)

III. Sozioökonomische Perspektive auf Transparenz und Nachvollziehbarkeit in algorithmischen Systemen

Um Transparenz und Nachvollziehbarkeit in algorithmischen Systemen aus sozioökonomischer Sicht besser beleuchten zu können, ist eine klarere Differenzierung notwendig. Aus sozioökonomischer Sicht müssen zur Bewertung von Transparenz und Nachvollziehbarkeit verschiedene Kontexte beleuchtet werden, so spielt die Art der beteiligten Akteure eine große Rolle. Akteure können dabei zu folgenden Gruppen gehören: staatliche Organisationen, Wirtschaftsunternehmen, intermediärer Sektor (wie Kirchen, Forschung und Vereine) sowie Bürgerinnen und Bürger.

Der Staat nimmt durch die Verwaltung öffentliche Aufgaben war, also Aufgaben, die explizit dem Gemeinwohl unterliegen. Durch eine ein Verfahren abschließende Entscheidung verpflichtet die Verwaltung die Betroffenen rechtsverbindlich zum Tun, Dulden oder Unterlassen, beispielsweise durch einen behördlichen Baubescheid oder eine Strafticket. Da sie entsprechend weitgehende Eingriffe in Rechte Dritter vornehmen kann, müssen Entscheidungen für die Öffentlichkeit transparent sein, um auch überprüfbar zu sein. Wirtschaftsunternehmen hingegen sind z.B. aus Gründen der Wettbewerbsfähigkeit, nicht in allen Bereichen verpflichtet, in ihren Handlungen transparent zu sein (Geschäftsgeheimnisse). Dies soll der Wertschöpfung und dem Erhalt des materiellen Wohlstands dienen. Im Rahmen der Rechtstaatlichkeit können Unternehmen zum Nachweis verpflichtet werden, dass z. B. einzelne Personen / Personengruppen keinen Schaden davontragen, benachteiligt oder ausgenutzt werden.

Datenbasierte Geschäftsmodelle, digitale Produkte und Serviceleistungen sind heute nicht mehr wegzudenken. Im Rahmen dieser wird die Individualisierung der Produkte immer mehr in den Mittelpunkt gestellt. Ob personalisierte Filmvorschläge, Sprachassistenten oder Gentests, sie haben alle das Ziel, für unterschiedliche Personen persönlich relevant zu sein. Diese Entwicklung hat dazu geführt, dass immer größere Mengen an Kundendaten anfallen und heute bereits in vielen algorithmischen Systemen verarbeitet werden. Es geht also bei der Nachvollziehbarkeit solcher algorithmischer Systeme auch um das Verständnis wie personenbezogene Daten verwendet und eingesetzt werden. Ebenfalls geht es darum zu prüfen, ob eingesetzte algorithmische Systeme sowohl korrekte als auch faire Entscheidungen treffen. Im Folgenden werden sowohl die Sichtweisen von Bürgerinnen und Bürgern als auch die Perspektiven von staatlichen Organisationen und Wirtschaftsunternehmen betrachtet.

These: Öffentlich zugängliche und vollständige Transparenz von algorithmischen Systemen ist nicht per se positiv.

Beschreibung: Eine pauschale Verpflichtung von Akteuren, ihre Daten, Modelle und Algorithmen öffentlich zugänglich und vollständig transparent zu machen, ist keine Antwort auf die hier aufgeworfene Frage nach dem richtigen Umgang mit Nachvollziehbarkeit. Es gilt abzuwägen, wie und für wen Transparenz oder Nachvollziehbarkeit hergestellt werden sollte. Die Befürchtung ist, dass eine umfängliche Transparenz für persönliche oder kriminelle Zwecke ausgenutzt werden könnte und in einem wirtschaftlichen oder gesellschaftlichen Schaden mündet.

Beispiel: Bei der Nutzung von Suchmaschinen könnten Kenntnisse des genauen Algorithmus dazu führen, dass bestimmte Produkte prominenter platziert werden, als für den Suchenden oder die Suchende relevant ist. Verzerrte Suchergebnisse und Benachteiligung kleinerer Webseiten führen zu wirtschaftlichen Schäden.

Umgang: Speziell für das genannte Beispiel könnten Gestaltende von Suchmaschinen verpflichtet werden, zwar grobe Einflussfaktoren für den Algorithmus zu nennen (z. B. optimale Formatierung oder unnatürliche Linkmuster), jedoch nicht das Modell bzw. die Entscheidungslogik dahinter (z. B. die genaue Gewichtung einer optimalen Formatierung oder die Erkennung unnatürlicher Linkmuster in der Indexierung) offenzulegen.

Allgemein betrachtet, könnte im ersten Schritt eine grobe Einteilung oder Klassifikation erfolgen, welche Akteure (z. B. staatliche Organe, Finanzinstitute, etc.) durch ihre Kernaufgaben oder Geschäftsmodelle besonders im Fokus der Aufklärung rund um Algorithmen-basierte Entscheidungen stehen oder ein besonders hohes Schadenspotenzial durch eine Offenlegung von Algorithmen haben. Im nächsten Schritt wäre es möglich, dass diejenigen, die zu mehr Transparenz verpflichtet werden, eine kontextspezifische Risikoabschätzungsanalyse ausführen. In

einem dritten Schritt muss festgelegt werden, was als Mindestmaß für Nachvollziehbarkeit gilt.[32]

These: Nachvollziehbarkeit ist mehr als technische Transparenz.

Beschreibung: Um Informationen nachvollziehen zu können, reicht es nicht aus, Informationen oder Daten transparent zu machen. Sie müssen in den jeweiligen Kontext gesetzt und erklärt werden. Die Nachvollziehbarkeit in algorithmischen Systemen ergibt sich somit aus einem interdisziplinären Ansatz. Neben den zugrundeliegenden Daten und algorithmischen Modellen, sollten auch nicht-technische Aspekte, wie der Algorithmus-Anwendungsfall, die wirtschaftliche Motivation dahinter oder die zugrundeliegenden ethischen Annahmen, ggf. auch Bias, berücksichtigt werden.

Beispiel: Bei der Vergabe von Krediten bzw. der Herleitung der Kreditwürdigkeit wird auf unterschiedliche Datenquellen zurückgegriffen. Hier werden unter anderem persönliche, sozioökonomische Daten mit herangezogen wie z. B. Wohnort, Sozialstatus und können so zu unterschiedlichen Ergebnissen führen. Nachvollziehbarkeit kann hergestellt werden, wenn eine Begründung vorliegt, wieso diese Daten verwendet werden und ob bzw. wie sie in die Auswertung einfließen. Mittels oben genannter Risikoabschätzungsanalyse könnten die Einflussfaktoren und die Motivation dahinter identifiziert werden.

Umgang: Zur Herstellung eines Mindestmaßes an Nachvollziehbarkeit könnten kontextspezifische Risikoabschätzungsanalysen eingeführt werden, wie sie bereits im Gesundheitsbereich üblich sind.[33] Hierbei sollte explizit auf technische, politische, wirtschaftliche und ethische Aspekte eingegangen werden. Beispiele wären:

- „Sind sozioökonomische Merkmale einzelner Personen, wie z.B. das Alter oder die Muttersprache, in den Daten enthalten und haben eine direkte Auswirkung auf die Ergebnisse der Berechnungen?“
- „Aus welchem Grund wurden diese Merkmale ausgewählt? Wie wurden die Abfragen integriert?“
- „Welche Möglichkeiten zur Änderung bzw. Löschung von Merkmalen werden Betroffenen angeboten?“

Die Antworten könnten folglich der Öffentlichkeit bereitgestellt werden, ohne dass zwangsweise auch die Modelle oder Algorithmen selbst veröffentlicht werden. Die damit verbundene Standardisierung, Erklärung in geeignetem Umfang und verständlicher Sprache würden so auch einem Laien die Nachvollziehbarkeit ermöglichen.

These: Standards dienen dazu, algorithmische Systeme vergleichbar zu machen.

Beschreibung: Gestaltende Personen algorithmischer Systeme benötigen Standards, damit sie einer Pflicht zur Nachvollziehbarkeit nachkommen können. Durch Entscheider und Entscheiderinnen definierte Standards können auch Betroffenen oder Prüfenden dienen, die beispielsweise die Nachvollziehbarkeit eines algorithmischen Systems in Frage stellen. Anhand dieser Kriterien kann besser und schneller evaluiert werden, ob es sich hier tatsächlich um ein nachvollziehbares System handelt oder nicht.

Beispiel: Unternehmen, die algorithmische Systeme einsetzen, sind verpflichtet entsprechende allgemeine Geschäftsbedingungen zu veröffentlichen und hier auch auf Datenschutz und -nutzung hinzuweisen. Aufgrund des Umfangs und der Komplexität dieser AGBs, ist es für Betroffene Personen trotz erfüllter Transparenz häufig nicht nachvollziehbar, wie und zu welchem Zweck personenbezogene Daten verwendet werden.

Durch eine Zertifizierung von Prozessen in der Planung, Entwicklung und im Einsatz können auch nicht vollständig transparente algorithmische Systeme bewertet werden.[34]

32 Prof. Dr. Zweig, K. (2019): Algorithmische Entscheidungen: Transparenz und Kontrolle; online verfügbar unter: https://www.kas.de/c/document_library/get_file?uuid=533ef913-e567-987d-54c3-1906395cdb81&groupId=252038 (letzter Abruf 14.06.2019)

33 Bundesinstitut für Risikobewertung (2010) Leitfaden für gesundheitliche Bewertungen; online verfügbar unter: https://mobil.bfr.bund.de/cm/350/leitfaden-fuer-gesundheitliche-bewertungen.pdf (letzter Abruf 14.06.2019)

34 KI Verband (2019): KI-Gütesiegel – AI Made in Germany; online verfügbar unter: https://ki-verband.de/ki-guetesiegel-ai-made-in-germany/ (letzter Abruf: 08.05.19)

Umgang: Ein möglicher Standard könnte beispielsweise eine Dokumentation in für die Betroffenen geeigneter Sprache und beschränktem Umfang sein, aus dem sofort ersichtlich wird, welche Daten von welchen Nutzergruppen im algorithmischen System verwendet werden. Weiterhin könnten zur Evaluation von Analysen Qualitätshandbücher aus der Statistik, wie das „Qualitätshandbuch der Statistischen Ämter des Bundes und der Länder"[35], verwendet werden.

IV. Ethisch-rechtliche Perspektive auf Transparenz und Nachvollziehbarkeit in algorithmischen Systemen

Aus ethisch-rechtlicher Sicht wird unter **Transparenz** eine allgemeine Informationsbereitstellung verstanden, ohne Anspruch darauf, dass diese Information aufbereitet ist. Durch eine zielgruppengerechte Aufbereitung können diese Informationen nachvollzogen werden. Hierbei ist nicht nur eine rein technische Nachvollziehbarkeit gemeint, sondern auch eine ethische, psychologische oder wirtschaftliche. Das heißt, dass Betroffene und Entscheidende nicht nur die technische Funktionsweise eines algorithmischen Systems verstehen sollen, sondern beispielsweise auch dessen Auswirkungen abwägen können müssen. Hierfür muss die Möglichkeit gegeben sein, Zusammenhänge zu hinterfragen, zu verstehen und zu überprüfen.

Zielgruppengerecht bedeutet, dass sich die Informationsbereitstellung der Zielgruppe anpassen muss, an die sie sich richtet. Allgemein muss auch hier zwischen Betroffenen, Gestaltenden, Entscheidenden und externen Prüfenden unterschieden werden. Denn während externe Prüfende mit vielen Informationen eine Nachvollziehbarkeit für sich schaffen können, benötigen Betroffene, denen es oftmals an technischer Kompetenz fehlt, eine allgemein verständliche Aufbereitung derselben Information. Eine einfache Offenlegung sämtlicher Informationen (um z.B. einer rechtlichen Vorgabe gerecht zu werden) würde bei ihnen das Gegenteil bewirken: Sie wären von der Menge an Informationen überfordert. Ein Übermaß an Informationen könnte sogar dazu führen, dass die wesentlichen Informationen nicht mehr wahrgenommen werden.

Aus ethisch-rechtlicher Sicht ermöglicht **Nachvollziehbarkeit** Personen einen Sachverhalt einzuschätzen und zu bewerten. Auch die DSGVO[36] stellt für die Transparenz einen Bezug zur Nachvollziehbarkeit her. Beispielsweise geht aus der DSGVO hervor, dass eine Information nur dann transparent ist, wenn sie präzise, leicht zugänglich und einfach zu verstehen ist. Wenn Kinder betroffen sind, muss die Information sogar in einer kindgerechten (und damit zielgruppengerechten) Sprache erfolgen. Ebenfalls muss die Information umso präziser und leichter zu verstehen sein, je größer die Anzahl an Akteuren und je höher die technologische Komplexität ist. Die DSGVO deckt indes viele Anwendungen algorithmischer Systeme nicht ab. Betroffene müssen zwar informiert werden, wenn und mit welchen Folgen eine vollautomatisierte Entscheidungsfindung oder eine Einstufung der Persönlichkeit für bestimmte Zwecke (Profiling) erfolgt. Dies öffnet jedoch viele Graubereiche, da die Abgrenzung zu lediglich vorbereitenden algorithmischen Systemen äußerst ungenau und einzelfallabhängig ist. Wenn beispielsweise ein algorithmisches System eingesetzt wird, um Bewerbungen auszusortieren, ist es fraglich, ob Personen diese vorbereitete Entscheidung hinterfragen. Das System entscheidet zwar nicht automatisch, wird aber eingesetzt, um den Entscheidungsprozess zu beschleunigen und zu vereinfachen. Zudem enthält die DSGVO keinerlei Regelungen zu den Auswirkungen von algorithmischen Systemen, wenn keine Personen, sondern möglicherweise „lediglich" Unternehmen oder aber der Wirtschaftsverkehr betroffen sind.

Vor diesem Hintergrund der ethisch-rechtlichen Betrachtung gibt es noch Handlungsbedarf hinsichtlich Transparenz und Nachvollziehbarkeit algorithmischer Systeme.

35 Destatis (2018): Qualitätshandbuch; online verfügbar unter: www.destatis.de/DE/Methoden/Qualitaet/Qualitaetshandbuch.pdf?__blob=publicationFile/ (letzter Abruf: 03.04.19)

36 Verordnung (EU) 2016/679 des europäischen Parlaments und des Rates zum Schutz natürlicher Personen bei der Verarbeitung personenbezogener Daten, zum freien Datenverkehr und zur Aufhebung der Richtlinie 95/46/EG (Datenschutz-Grundverordnung) (2016); online verfügbar unter: https://eur-lex.europa.eu/legal-content/DE/TXT/PDF/?uri=CELEX:32016R0679&from=DE (letzter Abruf: 23.05.2019)

These: Algorithmische Systeme sollten kontext-spezifisch und anhand des Diskriminierungs- und Schadenspotentials bewertet werden.

Beschreibung: Die Bewertung algorithmischer Systeme kann nicht anhand pauschaler Risikoklassen erfolgen, sondern muss immer auch den jeweiligen Einsatzkontext berücksichtigen, da dieser das jeweilige Diskriminierungs- und Schadenspotential maßgeblich beeinflusst.

Beispiel: Ein algorithmisches System, welches zum Risikoprofiling von Terroristen eingesetzt wird, weist ein hohes Diskriminierungs- und Schadenspotential auf.[37] Die Notwendigkeit gegenüber einer transparenten und nachvollziehbaren Aufbereitung wäre demnach höher als bei einem algorithmischen System, welches zur Steuerung einer Produktionsstrecke eingesetzt wird.

Umgang: Für eine effektive Umsetzung könnte eine Bewertung des Anwendungsfeldes anhand bestimmter Kriterien festgelegt werden. Bei der Ermittlung des Diskriminierungspotentials soll insbesondere eine Rolle spielen, ob Menschen direkt oder indirekt von dem Ergebnis eines algorithmischen Systems betroffen sind und ob das System Personen kategorisiert. Ebenso relevant ist die Abhängigkeit von dieser Entscheidung, beispielsweise wenn es keine Wechselmöglichkeit zu einem anderen Anbieter gibt.[38] Das Schadenspotential könnte sich zum einen aus den möglichen wirtschaftlichen, psychologischen und ökologischen Schäden, und zum anderen aus der Anzahl der Betroffenen zusammensetzen. Staatlich festgelegte Richtwerte, welche das Diskriminierungs- und Schadenspotential festlegen, könnten zur Orientierung dienen.

These: Eine allgemeine Pflicht, algorithmische Systeme zu jeder Zeit und in jedem Kontext vollständig nachvollziehbar aufbereiten zu müssen, führt zu einer Überregulierung, sodass Aufwand und Nutzen in keinem nützlichen Verhältnis stehen.

Beschreibung: Der Aufwand, der betrieben werden muss, um zu jeder Zeit, in jedem Kontext und für jegliche Zielgruppen vollkommene Nachvollziehbarkeit zu schaffen, ist immens.[39] Demnach muss bewertet werden ob, für wen und in welchem Ausmaß Nachvollziehbarkeit gewährleistet werden muss.

Beispiel: Bei einem algorithmischen System, das zur Steuerung einer Produktionsstrecke eingesetzt wird, ist das Diskriminierungs- und Schadenspotential gering, da zunächst keine Menschen betroffene sind. Bei Fehlentscheidungen würde der Schaden vorrangig dem Unternehmen zur Last fallen. Demnach würde der Aufwand, der betrieben werden müsste, um dieses algorithmische System zu jeder Zeit und in jedem Kontext nachvollziehbar zu gestalten, in keinem nützlichen Verhältnis stehen. Dem einsetzenden Unternehmen stünde es aber frei, vertraglich von dem Anbieter bzw. Entwickler einen höheren Grad an Transparenz oder Nachvollziehbarkeit zu fordern.

Umgang: Es sollte abgewogen werden, in wie weit Aufwand und Nutzen in einem nützlichen Verhältnis stehen. Die Notwendigkeit gegenüber einer nachvollziehbaren Aufbereitung soll über die Bewertung des Systems anhand des Diskriminierungs- und Schadenpotentials erfolgen. Zudem sollte klar definiert werden, für wen das System nachvollziehbar sein muss. Beispielsweise wäre es denkbar, dass algorithmische Systeme, die ein hohes Diskriminierungs- und Schadenspotential aufweisen, zu jeder Zeit sowohl für Betroffene als auch Entscheidende eine nachvollziehbare Aufbereitung leicht verfügbar machen müssten. Algorithmische Systeme, die ein geringes Diskriminierungs- und Schadenspotential aufweisen, müssten für Entscheidende, Gestaltende und externe Prüfende nachvollziehbar sein. Bei sicherheitskritischen Systemen sollte das System lediglich bei Prüfungen externer, staatlich legitimierter Institutionen eine nachvollziehbare Aufbereitung für Entscheidende und Prüfende vorweisen.

These: Algorithmische Systeme sollten immer transparent sein, wenn auch nicht für jeden.

Beschreibung: Damit im Falle einer Fehlentscheidung oder einer Prüfung Nachvollziehbarkeit erzielt werden kann, muss während des gesamten Entwicklungszyklus eines algorithmischen Systems für Transparenz gesorgt werden.

37 Bilger, A. (2018): Künstliche Intelligenz in der Verbrechensbekämpfung; online verfügbar unter: https://www.boell.de/de/2018/01/29/kuenstliche-intelligenz-der-verbrechensbekaempfung/ (Letzter Abruf 13.06.2019)
38 Prof. Dr. Zweig, K. (2019): Algorithmische Entscheidungen: Transparenz und Kontrolle; online verfügbar unter: https://www.kas.de/c/document_library/get_file?uuid=533ef913-e567-987d-54c3-1906395cdb81&groupId=252038 (letzter Abruf 14.06.2019)
39 DFKI (2017): Künstliche Intelligenz; online verfügbar unter: https://www.dfki.de/fileadmin/user_upload/import/9744_171012-KI-Gipfelpapier-online.pdf/ (letzter Abruf: 08.05.2019)

Für wen und in welchem Ausmaß diese Transparenz dann zur Nachvollziehbarkeit führen soll, ist vom Diskriminierungs- und Schadenspotential abhängig.

Beispiel: Ein Fitnesstracker, welcher das persönliche Bewegungsmuster aufzeichnet, hat auf den ersten Blick ein vergleichbar geringes Diskriminierung- und Schadenspotential. Die Auswertungen könnten jedoch bei Fehlentscheidungen möglicherweise zu einem gesundheitsschädlichen Verhalten oder später zu Abschlägen im Bonusprogramm von Krankenversicherungen führen. Spätestens in diesem Fall muss also auch die Nachvollziehbarkeit gewährleistet sein.

Umgang: Mit einer Dokumentation der Ziele, Daten, Methoden, Test- und Freigabeprozesse kann nicht nur eine höhere Qualitätssicherung, sondern auch eine umfassende Transparenz sichergestellt werden. Hierbei könnten Selbstdokumentationsmechanismen und Logs als Mindeststandard dienen. Wichtig wäre, dass im Falle einer Prüfung genug Daten gesammelt wurden, um eine post-hoc Analyse durchzuführen. Solch eine Daten-Transparenz hieße nicht, dass Betreiber jegliche Information jederzeit veröffentlichen müssten, jedoch auf Anfrage, beispielsweise bei einer Fehlentscheidung, relevante Information bereitstellen müssten. Je nach Diskriminierungs- und Schadenspotential würde entschieden werden, für wen und in welchem Ausmaß diese Transparenz gilt und für wen damit eine Nachvollziehbarkeit hergestellt werden muss. Beispielsweise müssten Betreiber eines algorithmischen Systems mit einem hohen Diskriminierungs- und Schadenspotential nicht nur für unternehmensinterne Transparenz, sondern auch für eine nach außen gehende Nachvollziehbarkeit sorgen. Die Bewertung des Diskriminierung- und des Schadenspotentials müsste kontinuierlich überprüft und ggf. geändert werden. Durch die Weiterentwicklung von algorithmischen Systemen entstehen Möglichkeiten, welche bei der Erstentwicklung oder beim ersten Einsatz noch nicht absehbar waren.

V. Ausblick

Vielen Menschen sind die Begriffe und die Auswirkungen rund um algorithmische Systeme noch nicht vertraut.[40] Forderungen nach informierter Einwilligung und digitaler Teilhabe können nur erfüllt werden, wenn Entscheidenden und betroffenen Menschen die Auswirkungen bekannt und bewusst sind. Es liegt in der Verantwortung von Fachleuten aufklärend tätig zu werden.

Die in diesem Denkimpuls vorgestellten Thesen zum Thema „Transparenz und Nachvollziehbarkeit algorithmischer Systeme“ erfordern Maßnahmen in ethisch-rechtlichen, sozioökonomischen und technologischen Bereichen. Im Denkimpuls werden Vorschläge für den konkreten Umgang genannt, die es nun zu diskutieren gilt. Wir empfehlen, existierende algorithmische Systeme vor diesem Hintergrund zu evaluieren und damit gleichzeitig die Wirksamkeit der vorgeschlagenen Maßnahmen zu prüfen.

Neben der Schwerpunktbetrachtung zum Thema „Transparenz und Nachvollziehbarkeit algorithmischer Systeme“ hat die Unterarbeitsgruppe Algorithmen-Monitoring in einem weiteren Denkimpuls bereits das Thema „Bias in algorithmischen Systemen“ bearbeitet und wird als nächstes das Thema „Verantwortung in algorithmischen Systemen“ behandeln.

40 D21-Digital-Index (2018 / 2019), Initiative D21; online verfügbar unter: https://initiatived21.de/publikationen/d21-digital-index-2018-2019/ (letzter Abruf: 18.06.2019)

Initiative D21

Kooperationspartner:

KPMG

Die Unterarbeitsgruppe Algorithmen-Monitoring

Algorithmische Systeme bergen ein immenses Potenzial, insbesondere kommt ihnen eine wachsende Bedeutung bei gesellschaftlichen Entwicklungen zu. Gleichzeitig entstehen eine zunehmende Komplexität und Intransparenz. Dies bringt steigende Herausforderungen und verschiedene Fragestellungen mit sich. Vor diesem Hintergrund hat die Initiative D21 Anfang 2018 eine Unterarbeitsgruppe (UAG) der AG Ethik zur Bearbeitung von Fragestellungen rund um das Thema „Algorithmen-Monitoring“ gegründet.

Die UAG Algorithmen-Monitoring diskutiert die relevanten Fragestellungen mit Expertinnen und Experten aus drei Perspektiven: ethisch-rechtlich, sozioökonomisch und technologisch. Dabei bezieht sich die technologische Perspektive auf die praktische Umsetzbarkeit eines Algorithmen-Monitorings und setzt sich mit den Bedingungen, Problemen und Möglichkeiten auseinander. Die sozioökonomische Perspektive arbeitet heraus, welche sozialen und ökonomischen Chancen und Herausforderungen durch die Anwendung von algorithmischen Systemen entstehen und wie man den Herausforderungen gegebenenfalls entgegenwirken kann. Die ethisch-rechtliche Perspektive behandelt die Erschließung einer rechtlichen Grundlage, welche die Regulierung algorithmischer Systeme ethisch vertretbar sichert.

Das Ziel der UAG Algorithmen-Monitoring besteht darin, für die drei Schwerpunktthemen „Bias in algorithmischen Systemen“, „Transparenz und Nachvollziehbarkeit algorithmischer Systeme“ und „Verantwortung in algorithmischen Systemen“ Thesen zu definieren und die Diskussionen zu Empfehlungen zusammenzufassen. Diese Empfehlungen sollen Vorschläge dazu enthalten, welche Regulierungen algorithmischer Systeme ethisch erforderlich sein könnten, wie sich diese gesellschaftlich und wirtschaftlich auswirken und wie sie technologisch umsetzbar wären.

ALGORITHMEN MONITORING

Impressum

Initiative D21 e.V.
Reinhardtstraße 38
10117 Berlin
www.InitiativeD21.de

Telefon: 030 5268722-50
kontakt@initiatived21.de

Download
initiatived21.de/publikationen/denkimpulse-zur-digitalen-ethik