Algo.Rules

Regeln für die Gestaltung
algorithmischer Systeme

Praxisleitfaden zu den Algo.Rules

Orientierungshilfen für Entwickler:innen und ihre Führungskräfte

iRights.Lab | BertelsmannStiftung

Inhaltsverzeichnis

Das Wichtigste im Überblick

Algorithmische Systeme unterstützen bei Entscheidungen, bereiten sie vor oder führen sie gar vollständig durch, und das zunehmend auch in Bereichen mit direktem Einfluss auf das Leben von Menschen: In Österreich analysiert man mittels Software beispielsweise die Daten Arbeitssuchender und teilt diese in Gruppen ein, damit Berater:innen der Arbeitsagentur anschließend gruppenspezifische Fördermaßnahmen empfehlen können. In Nordrhein-Westfalen analysieren algorithmische Systeme Kriminalitätsstatistiken und schlagen der Polizei vor, wo sie ihre nächste Streife fahren soll. Personalabteilungen setzen algorithmische Systeme ein, um Bewerbungen zu filtern, sortieren auf dieser Grundlage Kandidat:innen aus oder laden zu einem persönlichen Gespräch ein. Damit stellen solche Softwareanwendungen die Gesellschaft vor die Aufgabe, die Vorteile der Technologie für alle nutzbar zu machen - ohne dabei die potenziellen gesellschaftlichen und individuellen Risiken außer Acht zu lassen.

Viele Menschen, deren tagtägliche Arbeit es ist, diese oder ähnliche Software zu designen und zu entwickeln, sind von der positiven Gestaltungskraft der Technologie ebenfalls überzeugt. Doch etwa 60 Prozent der Entwickler:innen im Bereich Künstlicher Intelligenz (KI) haben Zweifel, ob die Softwareprodukte, an denen sie arbeiten, nicht auch negative Konsequenzen für Individuen oder Gesellschaft bedeuten können. Mehr als ein Viertel (27 Prozent) kündigt aus diesem Grund. Deshalb wünschen sich 78 Prozent praktische Hilfen, um über die gesellschaftlichen Auswirkungen ihrer Arbeit nachdenken zu können.[1] Das stellt IT- und Softwareunternehmen vor großen Herausforderungen, den ethischen Gestaltungsansprüchen der Gesellschaft und auch ihren Mitarbeiter:innen gerecht zu werden und geeignete Umsetzungshilfen dafür zu finden.

Damit das in der Praxis gelingt, braucht es Gestaltungsregeln für die Entwicklung und den Einsatz von algorithmischen Systemen. Deshalb entwickelten die Bertelsmann Stiftung im Projekt „Ethik der Algorithmen“ gemeinsam mit dem Think Tank iRights.Lab und unter Mitwirkung zahlreicher Expert:innen zunächst die Algo.Rules und erarbeiteten in einem breiten Konsortium Indikatoren und Kriterien für messbare „KI“-Ethik. Der vorliegende Praxisleitfaden und ein bald erscheinendes Impulspapier zur verzahnten Verantwortung von entwickelden und einsetzenden Organisationen bieten Orientierungshilfen zur Umsetzung der Regeln. Eine Handreichung für den öffentlichen Sektor erscheint in der zweiten Jahreshälfte 2020.

Bei den Algo.Rules handelt es sich um einen Katalog von neun Regeln, die beachtet werden müssen, um eine gesellschaftlich förderliche Gestaltung und den entsprechenden Einsatz algorithmischer Systeme zu ermöglichen und zu erleichtern. Diese Regeln sollen bereits bei der Entwicklung der Systeme beachtet und „by design“ implementiert werden. Sie richten sich an alle Personen, die einen signifikanten Einfluss auf die Entwicklung und den Einsatz algorithmischer Systeme haben. Dabei kommt den Entwickler:innen, Programierer:innen und Designer:innen eine besondere Rolle zu.

Entwickler:innen algorithmischer Systeme[2] und ihre Führungskräfte können wesentlich dazu beitragen, dass ethische Gestaltungsregeln wie die Algo.Rules in der Praxis beachtet werden. Sie können und sollten Software nicht nur nach technischen, sondern zugleich nach ethischen Gesichtspunkten gestalten. Auf technischer Ebene stellen sie die Weichen dafür, wie das System im späteren Einsatz funktioniert und wie es vorher definierte Ziele erreicht oder Probleme löst. Das spiegelt – oft auch unbewusst – Werteentscheidungen wider, z. B. darüber, wie transparent ein System gestaltet wird oder wie die Datenverarbeitung stattfindet. Entwickler:innen richten sich zumeist nach den Vorgaben ihrer Vorgesetzten und Auftraggeber:innen, die über die Ziele und Einsatzbereiche algorithmischer Systeme sowie über den Ressourceneinsatz bestimmen. Damit ist es die das System entwickelnde oder später einsetzende Organisation als Ganzes, die darüber bestimmt, welche Auswirkungen ein algorithmisches System auf Menschen und Gesellschaft hat.

1 Miller C, Coldicutt R. (2019) People, Power and Technology: The Tech Workers' View, London: Doteveryone. https://doteveryone.org.uk/report/workersview.
2 In Folge fassen wir die Berufsgruppen der Programmier:innen und Designer:innen unter der Gruppe der Entwickler:innen zusammen.

Arbeitsdefinition: Algorithmisches System
Ein Algorithmus ist eine eindeutige Handlungsvorschrift zur Lösung eines vorab definierten Problems. Ein algorithmisches System ist ein Softwaresystem einer oder mehrerer Algorithmen, um Daten zu erfassen, zu analysieren und Schlüsse zu ziehen, die zur Lösung eines vorher definierten Problems beitragen sollen. Diese Systeme werden auch algorithmenbasierte oder automatisierte Entscheidungssysteme genannt und in bestimmten Fällen Künstliche Intelligenz. Das System kann dabei selbstlernend sein oder vorprogrammierten Entscheidungsregeln folgen. Das algorithmische System umfasst aber auch seinen Einsatzrahmen: die Ziele und Rahmen, die Grundlage der Gestaltung des Systems sind; die Daten und Modelle, die in dieses eingespeist werden; die Deutung und Interpretation des Ergebnisses und die Ableitung einer Entscheidung durch Anwender:innen des Systems. Dieser sogenannte „soziotechnische Gesamtkontext" geht also weit über den reinen Programmiercode hinaus.

Es ist schwierig, genau festzustellen, wo der soziotechnische Gesamtkontext endet. Es hilft hierbei, ausgehend von Ziel und Wirkung zu denken: Welche Teile des Codes und des Kontexts tragen direkt oder indirekt zur Lösung des vorher definierten Problems bei? Welche haben Einfluss auf die Entscheidung, die vom System getroffen wird? Was trägt zu den (beabsichtigten und unbeabsichtigten) Wirkungen des Systems bei? Durch welche Prozesse wird die Entscheidung(sempfehlung) des Systems umgesetzt?

Dieser Leitfaden unterstützt diese Personengruppen dabei, die Algo.Rules in ihre Arbeitspraxis zu integrieren. Er zeigt anhand von Checklisten auf, welche Fragen sie sich stellen sollten, wenn sie algorithmische Systeme gestalten. Gleichzeitig kann der Leitfaden auch den Austausch zu ethischen Kriterien zwischen Entwickler:innen auf der einen und Führungskräften oder Anwender:innen auf der anderen Seite anleiten. Deshalb ist dieser Leitfaden auch für diese Gruppen relevant. Er stellt vor allem konkrete Maßnahmen vor, um die übergeordneten Regeln in der Praxis umzusetzen. Organisationen, die algorithmische Systeme entwickeln, können die Maßnahmen für ihre Zwecke anpassen und ergänzen oder sofort mit der Anwendung des Leitfadens beginnen.

Zu Anfang des Leitfadens stellen wir sechs zentrale Gründe vor, wieso sich die Arbeit mit den Algo.Rules für Entwickler:innen und das Unternehmen lohnt. Danach folgen Hinweise, wie man am besten mit diesem Leitfaden arbeitet. Darauf folgend kann anhand der sogenannten „Wirkungsrelevanz" geprüft werden, für welches algorithmisches System eine Implementierung der Algo.Rules besonders geeignet ist. Die weiteren Kapitel widmen sich den einzelnen Algo.Rules. Dafür leitet zunächst die Algo.Rule selbst das Kapitel ein. Eine weitere Konkretisierung verdeutlicht die Tragweite der Regel und gibt Hinweise zur Definition bestimmter Begrifflichkeiten. Nutzer:innen des Leitfadens können anhand der Erfolgsfaktoren erkennen, wie die Regel umzusetzen wäre (z. B. frühzeitig, inklusiv oder unter Einbezug weiterer Gruppen). Darauf folgen Orientierungsfragen, die sich Entwickler:innen einzeln oder im Team stellen können, um die Algo.Rules umzusetzen. Der Leitfaden ist modular aufgebaut, sodass jede Regel für sich genommen oder im Gesamtkomplex eine Basis bietet, um sie zu behandeln und zu verwenden.

WARUM?

Sechs Gründe, die Algo.Rules umzusetzen

Neben der Wahrnehmung gesellschaftlicher Verantwortung bringt eine Beachtung der Algo.Rules auch große Vorteile für die entwickelnden und einsetzenden Organisationen mit sich. Zu diesen potenziellen Vorteilen gehören:

1 Prävention:

Durch eine Verankerung der Algo.Rules bereits im Entwicklungsprozess („Algo.Rules by Design“) werden vermeintliche Fehlentwicklungen vermieden, die möglicherweise (auch finanzielle) Schäden für Organisationen, Personen oder die Gesellschaft bedeutet hätten.

2 Effektivität:

Algorithmische Systeme können ihre Aufgaben besser erfüllen und die ihnen gegebenen Ziele besser erreichen (z. B. durch die Algo.Rule „Nachvollziebarkeit herstellen“). Das kann aufseiten der entwickelnden wie auch einsetzenden Organisationen zu Einsparungspotenzialen führen.

3 Laufende Verbesserung:

In den Algo.Rules sind Mechanismen vorgesehen, die eine laufende Überprüfung der Funktionsweise des algorithmischen Systems erlauben, um Fehler frühzeitig zu erkennen und zu beheben (z. B. durch die Algo.Rule „Wirkung überprüfen“).

4 Attraktivität für Kund:innen:

Kund:innen haben einen Bedarf nach der Beachtung gesellschaftlich-ethischer Prinzipien wie den Algo.Rules. Sie werden algorithmische Systeme, die die Algo.Rules beachten, eher kaufen und nutzen. So kann ein Marktvorteil gegenüber anderen Anbieter:innen gewonnen werden.

5 Vertrauen der Nutzer:innen:

Die späteren Nutzer:innen und Betroffenen wollen mit einem vertrauenswürdigen algorithmischen System zu tun haben. Die Algo.Rules können die Akzeptanz eines Systems erhöhen (z. B. Kennzeichnung durchführen, Beschwerden ermöglichen).

6 Attraktivität für Entwickler:innen:

Entwickler:innen werden sich zunehmend ihrer gesellschaftlichen Verantwortung bewusst und verlangen schon heute von potenziellen Arbeitgebern, sich auch über den gesetzlichen Rahmen hinaus an ethisch-gesellschaftlichen Standards zu orientieren. Sie fordern diesen Anspruch zunehmend bei der Arbeitssuche und der Auswahl von Arbeitgebern ein. Mit den Algo.Rules können Organisationen ihre Attraktivität im Wettbewerb um Fachkräfte stärken.

WIE? Die Arbeit mit dem Leitfaden

Der Leitfaden ist eine Einladung zum Nachdenken und Diskutieren. Er liefert bewusst keinen genauen, feingranularen Umsetzungspfad für die Algo.Rules, sondern stellt wichtige grundsätzliche Punkte und Fragen zusammen. Die Umsetzung der Algo.Rules hängt vom jeweiligen Fall ab: Von den beteiligten Individuen, Teams und Organisationen, vom Einsatzkontext und -ziel, von den verfügbaren Ressourcen und Kompetenzen. Personen und Teams, die sich mit den ethisch-gesellschaftlichen Auswirkungen ihrer Arbeit und ihres aktuellen Projekts, Sprints oder Produkts beschäftigen wollen, können hier einen Start- und Anhaltspunkt finden. Ein Vorgehen kann beispielsweise wie folgt angelegt werden:

- **Früh beginnen:**
  Die Algo.Rules sollen von Anfang an in die Gestaltung algorithmischer Systeme einbezogen werden. Deshalb gibt es keinen Zeitpunkt, der zu früh ist, um diesen Leitfaden in die Hand zu nehmen.

- **Relevanz aufzeigen:**
  Zunächst empfiehlt es sich, innerhalb des Teams oder der Organisation die Algo.Rules bekannt zu machen. Überlegen Sie sich gemeinsam, warum gerade Ihre Arbeit gesellschaftliche Relevanz hat. Zeigen Sie auf, dass die Beachtung der Algo.Rules nicht (nur) Mehrarbeit bedeutet, sondern Sie und Ihre Organisation davon profitieren können.

- **Kritisches Mindset etablieren:**
  Ermutigen Sie die kritische Grundhaltung, dass Ihr algorithmisches System Fehler haben wird und negative Auswirkungen auf die Gesellschaft keine Ausnahme darstellen. Kein System ist perfekt, Risikominimierung ist ein laufender Prozess.

- **Überblick schaffen:**
  Zu Beginn eines neuen Projekts kann es helfen, einen Überblick über die für diesen Fall möglichen spezifischen ethisch-gesellschaftlichen Auswirkungen zu gewinnen. Dafür kann eine Ermittlung der Wirkungsrelevanz durchgeführt werden.

- **Ansatzpunkt identifizieren:**
  Identifizieren Sie im Rahmen der Ermittlung der Wirkungsrelevanz einen bestimmten Teilbereich der Algo.Rules, der sich für die Umsetzung in Ihrem Projekt besonders eignet, z. B. weil gerade eine Kompetenzlücke im Team besteht oder weil Sie sich unter dieser Algo.Rule am einfachsten etwas Konkretes für Ihr Projekt vorstellen können. Schauen Sie, wo Sie „Quick Wins“ erreichen können.

- **Koordinieren:**
  Die Umsetzung jeder Algo.Rule erfordert ein Zusammenspiel der an der Gestaltung des algorithmischen Systems beteiligten Personen. Besprechen Sie, wer welche Rolle übernehmen kann, sprechen Sie mit Ihrem Team, relevanten Mitarbeiter:innen aus anderen Abteilungen, Führungskräften, Anwender:innen und anderen Beteiligten.

- **Fürs eigene Projekt personalisieren:**
  Sammeln Sie Inspiration in diesem Leitfaden. Überlegen Sie, auch gemeinsam im Team und mit anderen Beteiligten, welcher Weg für Sie passend ist. Möglicherweise beinhaltet das die relativ unkomplizierte Durchführung eines Workshops, die Dokumentation von gewissen Prozessen oder Anpassung von Code und Design. Möglicherweise werden Sie aber auf Herausforderungen stoßen, die nur mit aufwendigeren Veränderungen zu bewältigen sind.

- **Nächste Schritte planen:**
  Die Algo.Rules sind interdependent aufgebaut. Sie werden bei der Umsetzung einer Algo.Rule merken, dass auch Aspekte anderer Algo.Rules relevant sind. Eine schrittweise Umsetzung kann sich so auch auf natürliche Weise entwickeln. Planen Sie Ihre weiteren Schritte.

- **Den Blick fürs Ganze behalten:**
  Behalten Sie deshalb die Algo.Rules als Ganzes im Blick. Nur als Ganzes können Sie Ihrem Ziel gerecht werden, die Chancen algorithmischer Systeme zu nutzen und Risiken zu verringern.

WAS?

Orientierungshilfen zur Anwendung der Algo.Rules

Vorab: Wirkungsrelevanz ermitteln

Die Algo.Rules erheben nicht den Anspruch, auf alle algorithmischen Systeme angewendet zu werden. Die Arbeit von Entwickler:innen hat gesellschaftliche Auswirkungen – aber nicht immer sind diese Auswirkungen gleichbedeutend. Im Fokus der Algo.Rules stehen jene algorithmischen Systeme, **die einen direkten oder mittelbaren, immer jedoch signifikanten Einfluss auf das Leben der Menschen oder die Gesellschaft haben.** Dies ist insbesondere dort der Fall, wo algorithmische Systeme in menschliche Entscheidungsprozesse eingebunden sind und so beispielsweise über die Zuteilung von Geld, Gütern, Chancen oder Freiheiten mitbestimmen.

Um zu bestimmen, ob die Algo.Rules für ein System relevant sind, sollte vor der Gestaltung des Systems eine Einschätzung seiner Wirkung getroffen werden. Je stärker der mögliche Einfluss eines algorithmischen Systems auf das Leben der Menschen oder die Gesellschaft ist, desto gründlicher sollte die Umsetzung der Algo.Rules und der Orientierungsfragen geprüft werden. Der Grund ist, dass Systeme mit einem niedrigeren Einfluss, wie z. B. eine automatisierte Kühlung, nicht so aufwendig geprüft werden müssen wie beispielsweise Systeme, die zur Krankheitsdiagnose in Krankenhäusern eingesetzt werden.

Konkretisierung:

Die Wirkungsrelevanz liefert eine Einordnung des entsprechenden algorithmischen Systems. Abhängig davon ist im Anschluss zu bestimmen, welchem Schutzlevel die Gestaltung des algorithmischen Systems genügen muss – also welche der Algo.Rules in welchem Umfang und mit Blick auf welche der Orientierungsfragen umzusetzen sind. Um die Wirkungsrelevanz zu bestimmten, müssen das algorithmische System und die Umgebung, in der es eingesetzt wird, betrachtet werden. Entscheidende Faktoren sind dabei die Schwere des potenziellen Schadens und die Abhängigkeit der Betroffenen von dem entsprechenden algorithmischen System. In den Orientierungsfragen wird weiter präzisiert, was dabei zu beachten ist.

Erfolgsfaktoren für entwickelnde Organisationen:

- Wirkungsrelevanz frühzeitig, z. B. zu Beginn der Produktentwicklung, ermitteln
- Soziotechnischen Gesamtkontext des algorithmischen Systems bestimmen
- Externe Expert:innen, insbesondere Sozialwissenschaftler:innen und/oder Sozioinformatiker:innen, sowie Beteiligte und Betroffene einbeziehen
- Unmittelbar und mittelbar vom algorithmischen System betroffene Personen(gruppen) identifizieren

Orientierungsfragen für die Ermittlung der Wirkungsrelevanz:

Bestimmung des potenziellen Schadens des algorithmischen Systems (gemäß Krafft und Zweig 2019)

- ▢ Wie groß sind die potenziellen Auswirkungen auf die individuelle Ausübung der Grundrechte, Gleichheit und Gerechtigkeit? Wie groß ist das Schadenspotenzial – auch bei möglichen Fehlern – des algorithmischen Systems?
- ▢ Wie viele Menschen sind von einem algorithmischen System betroffen? Über wie viele Menschen wird mithilfe des algorithmischen Systems entschieden?
- ▢ Gibt es, unabhängig von den direkten Auswirkungen, ein mögliches Risiko für die Gesellschaft als Ganzes?
- ▢ Wie hoch ist die Wahrscheinlichkeit, dass diese Schäden eintreten?

Bestimmung der Abhängigkeit der Betroffenen vom algorithmischen System (gemäß Krafft und Zweig 2019)

- ▢ Wird der Output des algorithmischen Systems zusätzlich von einem Menschen überprüft, bevor er zu einer Entscheidung mit Auswirkungen für Betroffene führt? Wie groß ist die menschliche Kontrolle über das System?
- ▢ Besteht die Möglichkeit für Betroffene, sich der algorithmenbasierten Entscheidung zu entziehen, ohne negative Konsequenzen zu befürchten?
- ▢ Ist die algorithmenbasierte Entscheidung umkehrbar? Können mögliche Fehler korrigiert und Schäden behoben werden?

Bestimmung der Anforderungen für die Gestaltung des algorithmischen Systems

- ▢ Welches sind die sich aus der Wirkungsrelevanz ergebenden Anforderungen für die Gestaltung des algorithmischen Systems?
- ▢ Welche konkreten Maßnahmen sind durchzuführen, um die ermittelten negativen Auswirkungen zu minimieren? Welche Maßnahmen sind dabei sinnvollerweise prioritär umzusetzen?

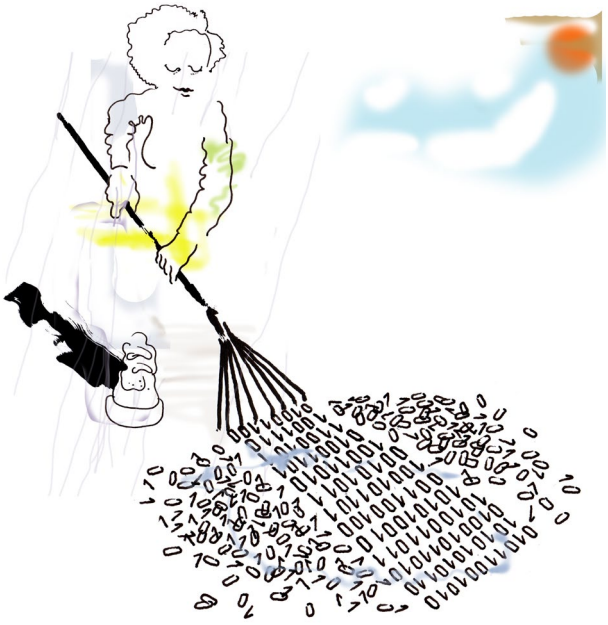
Weiterführende Links zu Forschung, Ansätzen und Impulsen:

Der Ansatz der Risikomatrix von Krafft und Zweig (2019), auf der die hier vorgestellte Ermittlung der Wirkungsrelevanz basiert:
https://t1p.de/Systeme-Link-1 https://t1p.de/Systeme-Link-2

Verfahren zur Berechnung des Teilhabewirkungspotenzials algorithmischer Entscheidungssysteme:
https://t1p.de/Systeme-Link-3

Workshopkonzept zur Identifizierung ethischer Fragen der Arbeit an datengetriebenen Projekten:
https://t1p.de/Systeme-Link-4

Visuelles Verfahren zum Schaffen eines Überblicks über den soziotechnischen Kontext eines algorithmischen Systems:
https://t1p.de/Systeme-Link-5



1 Kompetenz aufbauen

Die Funktionsweise und die möglichen Auswirkungen eines algorithmischen Systems müssen verstanden werden.

Diejenigen, die algorithmische Systeme entwickeln, betreiben und/oder über ihren Einsatz entscheiden, müssen über die erforderliche Fachkompetenz und ein entsprechend abgestuftes Verständnis der Funktionsweisen und potenziellen Auswirkungen der Technologie verfügen. Das Teilen von individuellem und institutionellem Wissen sowie der interdisziplinäre Austausch zwischen den Aufgabenbereichen sind dafür ebenso zentral wie qualifizierende Maßnahmen. Diese sind in die Ausbildung bzw. das Onboarding neuer Mitarbeiter:innen zu integrieren. Der interdisziplinäre Austausch sollte verstetigt werden und für andere Interessierte und/oder betroffene Akteur:innen offenstehen.

Konkretisierung:

Der Aufbau von Kompetenz umfasst je nach Zielgruppe Basiskompetenzen, wie kritisches Denken und Diskussionfähigkeit, sowie Fachkompetenzen, wie Programmierfähigkeiten oder Kenntnisse des gesellschaftlichen Anwendungsfeldes als auch eine grundlegende Sensibilisierung für die gesellschaftlich-ethischen Dimensionen der Digitalisierung und der konkreten Arbeit. Dabei ist es für Entwickler:innen besonders wichtig, sich solche Kompetenzen anzueignen, die in ihrer Ausbildung eher zu kurz kommen. Dazu zählen insbesondere soziale, ethische und rechtliche Aspekte. Wiederum andere Personen, die beispielsweise das System in der Praxis einsetzen, brauchen ein technisches Grundverständnis über algorithmische Systeme. Nur so ist es möglich, dass sich alle, die an algorithmischen Systemen mitwirken, gegenseitig verstehen und damit eine gemeinsame Sprache gefunden haben.

Die Bedeutung dieser Algo.Rule geht deshalb weit über das einzelne Projekt hinaus. Denn damit die an der Entwicklung und dem Einsatz involvierten Personen über die unterschiedlichen Felder hinweg Kompetenzen aufbauen können, müssen langfristige Maßnahmen getroffen werden. Diese sollten in der gesamten Organisation und über Teams hinweg durchgeführt werden. Weiterhin reicht die Frage des Kompetenzaufbaus auch über einzelne Institutionen hinweg bis in die Ausbildung von Entwickler:innen oder anderen Beteiligten.

Erfolgsfaktoren für entwickelnde Organisationen:

- Kompetenzbedarfe identifizieren
- Raum für Kompetenzaufbau und Austausch schaffen sowie Ressourcen bereitstellen
- Kompetenzaufbau projektübergreifend, interdisziplinär und zielgruppengerecht durchführen
- Kompetenzanforderungen auf dem aktuellen Stand halten und Anpassungsbedarf regelmäßig überprüfen
- Maßnahmen zum Kompetenzaufbau laufend überprüfen und an sich ändernde Umstände und Technologien anpassen
- Expert:innen, Anwender:innen und Betroffene zum Kompetenzaufbau konsultieren
- Kompetenzanforderungen auch für Anwender:innen und Betroffene antizipieren und, falls möglich, gegenüber Kund:innen artikulieren

Orientierungsfragen für einen erfolgreichen Kompetenzaufbau:

Schaffung des organisatorischen Rahmens für einen nachhaltigen Kompetenzaufbau

- ▢ Wäre eine zentrale Stelle, die den Kompetenzaufbau innerhalb der Organisation koordiniert, sinnvoll und wie sollte diese aufgebaut sein?

Vorbereitung des Kompetenzaufbaus

- ▢ Wer sind die an der Gestaltung des algorithmischen Systems beteiligten Personen(gruppen)? Was sind die individuellen Erfordernisse für den Kompetenzaufbau dieser beteiligten Gruppen?
- ▢ Welche Lücken gibt es in den Kompetenzen der beteiligten Personen?
- ▢ Werden beim Kompetenzaufbau unterschiedliche Dimensionen (technische Kompetenz, anwendungsorientierte Kompetenz, gesellschaftlich-kulturelle Kompetenz, juristische Kompetenz) beachtet und abgebildet?
- ▢ Ist der Kompetenzaufbau zielgruppenspezifisch hinsichtlich unterschiedlicher Rollen (Frontend, Backend, UX/UR, Teamleitung etc.) individualisiert?
- ▢ Was sind relevante technische Entwicklungen und wie werden Informationen über diese gesammelt?

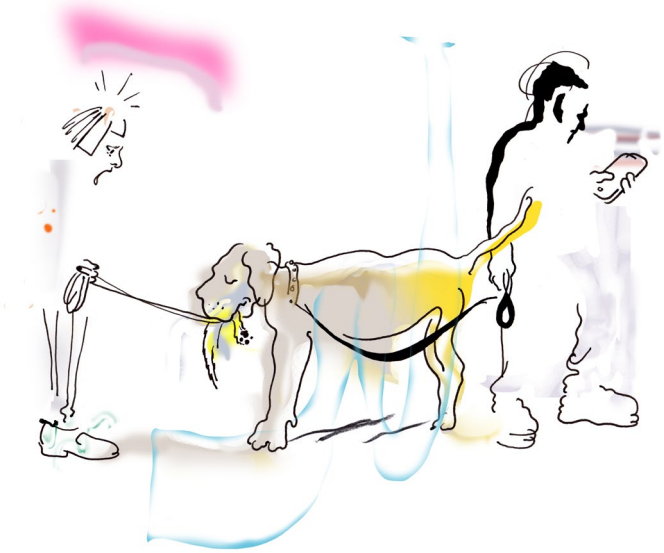
Maßnahmen für den Kompetenzaufbau

- ▢ Welche wichtigen Begriffe sollten organisationsweit einheitlich verwendet werden? Wurde ein entsprechendes Glossar oder Wiki erstellt?
- ▢ Werden in einem Wissensspeicher projektübergreifend relevante Erfahrungen gesammelt und zur Verfügung gestellt werden?
- ▢ Werden Basiskompetenzen, wie Diskursfähigkeit, Reflexionsfähigkeit, interdisziplinäres Arbeiten, Kommunikationsfähigkeit und hypothesengeleitetes Arbeiten, laufend und zielgruppenübergreifend gefördert?
- ▢ Wie gestaltet sich der interdisziplinäre Austausch innerhalb der Organisation? Werden beispielsweise feste Workshops, offene Diskussionsabende, Umfragen, Meinungsbildungsprozesse, projektspezifische Diskussionssprints, Show-and-Tell-Formate oder weitere Methoden zu diesem Zwecke genutzt?
- ▢ Wird darüber eine Sensibilisierung für die Perspektiven anderer erreicht?
- ▢ Werden in Pitches an Führungskräfte, sowohl intern als auch gegenüber Kund:innen, gesellschaftlich-ethische Aspekte integriert?
- ▢ Wie ist das Onboarding für neue Mitarbeiter:innen gestaltet?

Weiterführende Links zu Forschung, Ansätzen und Impulsen:

Überblick über verschiedene Dimensionen digitaler Kompetenzen und Bildung:
https://t1p.de/Kompetenz-link-1

Studie über die Position eines Chief Compliance Officer, der den Kompetenzaufbau in einem Unternehmen koordinieren kann:
https://t1p.de/Kompetenz-link-2



2 Verantwortung definieren

Für die Auswirkungen des Einsatzes eines algorithmischen Systems muss stets eine natürliche oder juristische Person verantwortlich sein.

Die Verantwortung bedarf einer eindeutigen Zuteilung. Die damit verbundenen Aufgaben müssen der zuständigen Person bewusst sein. Dies gilt auch für geteilte Verantwortlichkeiten bei mehreren Personen oder Organisationen. Die Zuteilung muss lückenlos sowie nach innen und außen transparent dokumentiert werden. Die Verantwortung darf weder auf das algorithmische System noch auf die Anwender:innen oder betroffenen Personen abgewälzt werden.

Konkretisierung:

Auch beim Einsatz (komplexer) algorithmischer Systeme bleibt der Mensch verantwortlich, weil die Ziele algorithmischer Systeme und ihre technischen Funktionen von Menschen gestaltet werden. Verantwortung umfasst dabei zwei Dimensionen: Zum einen sind Personen verantwortlich für die Umsetzung von Maßnahmen und die Gestaltung des algorithmischen Systems. Zum anderen gehört rechtliche Verantwortlichkeit dazu: Sie wirkt zumeist im Nachhinein und adressiert Fragen der Haftung.

Was genau die Verantwortlichkeiten in beiden Dimensionen sind, muss konkretisiert und eingegrenzt werden. So muss bereits während der Gestaltung des algorithmischen Systems und auch während des Einsatzes klar sein, wer oder welche Organisation inwiefern und für welche Auswirkungen des Systems verantwortlich ist. Klar definierte Verantwortlichkeiten schärfen das eigene Rollenverständnis. Sie verhindern, dass eine diffuse Verantwortung auf das technische System selbst übertragen wird.

Erfolgsfaktoren für entwickelnde Organisationen:

- Entwicklungskontext und die Anwendung des algorithmischen Systems mit allen Wirkzusammenhängen umfassend aufbereiten
- Verantwortung trotz der Mitarbeit vieler Menschen und Organisationen klar definieren
- Lieferketten und Unterauftragsnehmer angemessen in die Verantwortung nehmen
- Bei möglichen Fehlern Verantwortung zur Beseitigung und zukünftigen Vermeidung von Schäden wahrnehmen
- Verantwortlichkeiten regelmäßig überprüfen und anpassen

Orientierungsfragen für definierte Verantwortlichkeiten:

Zuteilung der Verantwortlichkeiten

- ▢ Wurde einer juristischen Person die Verantwortung für die Fertigstellung des algorithmischen Systems übertragen und klärt die Organisation ihre Verantwortlichkeit intern selbst?
- ▢ Wurde eine zentrale Verantwortungsperson benannt, die das Projekt überblickt und alle Gestaltungsprozesse nachvollziehen und mitentscheiden kann und die hierzu auch von der Geschäftsführung ermächtigt ist?
- ▢ Wurde ein System abgestufter Freigabeverantwortung eingerichtet, bei der für jeden Gestaltungsschritt jeweils eine Person verantwortlich ist und die Verantwortung über die Freigabe bzw. Abnahme übertragen wird?
- ▢ Oder wurde ein System verteilter Verantwortung, wie z. B. eine Verantwortungskette oder rollenspezifische Verantwortungsbereiche, eingerichtet?

Information der verantwortlichen Personen

- ▢ Welches sind die Rollen, Rechte und Pflichten der Verantwortlichen?
- ▢ Was bedeutet das für ihre Arbeit? Welche Maßnahmen treffen die Verantwortlichen für sich und ihren Arbeitsbereich?

Dokumentation der Verantwortlichkeiten

- ▢ Sind die Verantwortlichkeiten in der Dokumentation klar geregelt?
- ▢ Ist eine Ansprechperson (auch für Externe einsehbar) benannt?
- ▢ Kann die Dokumentation auf Anfrage von Kund:innen oder Betroffenen sowie legitimierten Aufsichtsbehörden zur Verfügung gestellt werden?

Weiterführende Links zu Forschung, Ansätzen und Impulsen:

Arbeitspapier zur Verkettung von Verantwortlichkeiten und der Fehleranalyse bei algorithmischen Systemen:
https://t1p.de/Verantwortung-link-1

Forschungspapier zur verteilten moralischen Verantwortlichkeit bei Teams, u.a. in der Softwarebranche:
https://t1p.de/Verantwortung-link-2

Überblicksartikel zur RACI- bzw. CAIRO-Methode zur Erstellung einer Verantwortungsmatrix:
https://t1p.de/Verantwortung-link-3

3 Ziele und erwartete Wirkung dokumentieren

Die Ziele und die erwartete Wirkung des Einsatzes eines algorithmischen Systems müssen vor dessen Einsatz dokumentiert und abgewogen werden.

Die Ziele des algorithmischen Systems müssen klar definiert und Informationen zu dessen Einsatz dokumentiert werden. Dazu zählen etwa die zugrunde liegenden Daten und Berechnungsmodelle. Vor dem Einsatz des algorithmischen Systems muss eine dokumentierte Folgenabschätzung durchgeführt werden. Insbesondere bei lernenden Systemen und in dynamischen Einsatzfeldern mit viel Veränderung ist die Folgenabschätzung in regelmäßigen Abständen zu wiederholen. Hierbei sind Risiken für Diskriminierungen und weitere für das Individuum und das Gemeinwohl berührende Folgen im Blick zu behalten. Werteabwägungen bei der Zielsetzung und dem Einsatz von algorithmischen Systemen müssen festgehalten werden.

Konkretisierung:

Durch den Einsatz algorithmischer Systeme sollen mehr oder weniger klar eingegrenzte Problemstellungen gelöst und festgelegte Organisations- oder Projektziele erreicht werden. Es ist für die entwickelnde oder einsetzende Organisation sinnvoll, den dahinterliegenden Zielfindungsprozess zu dokumentieren. Ein transparenter Prozess soll dabei helfen, mögliche Zielkonflikte und -abwägungen, Wertegrundlagen, Rahmenbedingungen und Nicht-Ziele (zu vermeidende Ergebnisse) vor und während der Entwicklung aufzudecken und schriftlich festzuhalten.

Eine Folgenabschätzung soll auf Basis der festgelegten Ziele potenzielle Auswirkungen eines Einsatzes konkret aufzeigen und eine Risiko- oder Schadensabwägung ermöglichen. Sie kann mit den festgelegten Zielen kontrastiert werden, um zu überprüfen, ob diese erreicht wurden und ob sich neben den beabsichtigten Wirkungen auch unbeabsichtigte und/oder negative (Langzeit)Wirkungen ergeben haben. Zur Analyse gehört dabei der gesamte soziotechnische Kontext des algorithmischen Systems, also auch die Daten – Stammdaten, Eingabe- und Ausgabedaten sowie Trainingsdaten – und Modelle, die in das algorithmische System einfließen.

Erfolgsfaktoren für entwickelnde Organisationen:

- Widersprüchliche Ziele und ethische Dilemmata sichtbar machen
- Folgenabschätzungsprozess und damit verbundene Ressourcen frühzeitig einplanen
- Messbare Indikatoren auswählen und klar definieren, damit die beabsichtigte Wirkung und Erreichung der Ziele in der Praxis getestet werden können
- Erwartete Wirkung möglichst ganzheitlich erfassen
- Für die Grenzen der Folgenabschätzung sensibilisieren und auf unbekannte Wirkungen vorbereitet sein
- Komplexe kausale Zusammenhänge verständlich aufbereiten

Orientierungsfragen für dokumentierte Ziele und Wirkungen:

Bestimmung der Ziele des algorithmischen Systems

- ▢ Wurden die Ziele und Nicht-Ziele des algorithmischen Systems definiert?
- ▢ Was sind mögliche Zielkonflikte?
- ▢ Ist ein algorithmisches System das richtige Mittel, um diese Ziele zu erreichen?
- ▢ Was ist Zielvorgabe für den Entwicklungsprozess des algorithmischen Systems (definition of ready, definition of done)?
- ▢ Sind die verschiedenen Ziele konsistent und insgesamt vollständig?
- ▢ Was sind die (institutionellen, finanziellen etc.) Anreize, nach denen die Ziele bestimmt wurden, und wie werden diese reflektiert und dokumentiert?
- ▢ Sollte es sich um eine Auftragsarbeit handeln, wie werden die Ziele mit den Kund:innen reflektiert?
- ▢ Beinhaltet die Angebotsbeschreibung diese Ziele und mögliche Konflikte?

Ermittlung der Betroffenen des algorithmischen Systems

- ▢ Sind alle direkt und indirekt an der Gestaltung des Systems beteiligten und vom Einsatz des Systems betroffenen Personen(gruppen) identifiziert?
- ▢ Werden dabei diverse Meinungen von Expert:innen, Gesellschaftsvertreter:innen oder Betroffenen aufgenommen und beachtet?

Abschätzung der Folgen des algorithmischen Systems

- ▢ Was sind die Auswirkungen des Systems auf Güter, Personen und Prozesse?
- ▢ Was sind die Konsequenzen für unterschiedliche Personen(gruppen), wie Anwender:innen des algorithmischen Systems, direkt und indirekt Betroffene sowie die Gesellschaft insgesamt?
- ▢ Welche Szenarien/User Stories für unterschiedliche Situationen gibt es? Welche möglichen Fehlerquellen ergeben sich daraus?
- ▢ Wie sind die soziotechnischen Risiken des algorithmischen Systems strukturiert?
- ▢ Sind die Grenzen der Gültigkeit der Folgenabschätzung reflektiert und transparent dokumentiert?
- ▢ Wie werden Betroffene und externe Expert:innen an der Entwicklung der Folgenabschätzung beteiligt?
- ▢ Wurden der Prozess der Folgenabschätzung und andere Inspektionstechniken geprüft?
- ▢ Wie könnte ein automatisiertes Testverfahren aussehen, das mittels Simulationen maschinelle Folgenabschätzungen durchführt?
- ▢ Was bedeuten die Ergebnisse der Folgenabschätzung für die Maßnahmen von Algo.Rule 8 (Wirkung überprüfen)?

Überprüfung der technischen Elemente

- ▢ Wurden das Trainingsdatenset, das Inputdesign (Sensoren, User Interface), die Inputdaten, die angewendeten Methoden und Metakriterien sowie das Output Design auf mögliche Diskriminierungen überprüft?
- ▢ Können möglicherweise selbstverstärkende Prozesse im Einsatz auftreten?
- ▢ Welche Modelle werden verwendet und was sind ihr Nutzen, ihre Risiken, Durchführungsmöglichkeiten und Unsicherheiten?

Weiterführende Links zu Forschung, Ansätzen und Impulsen:

Prototyp für die einheitliche Beschreibung und Bewertung von Datensets:
https://t1p.de/Ziele-Link-1

Datenschutzfolgeabschätzung als Inspiration für Folgenabschätzung:
https://t1p.de/Ziele-Link-2

Ansatz eines KI-Ethiklabels als Inspiration für Folgenabschätzung:
https://t1p.de/Ziele-Link-3

4 Sicherheit gewährleisten

Die Sicherheit eines algorithmischen Systems muss vor dessen Einsatz getestet und fortlaufend gewährleistet werden.

Zuverlässigkeit und Robustheit eines algorithmischen Systems und der zugrunde liegenden Daten gegenüber Angriffen, Zugriffen und Manipulationen sind unbedingt zu gewährleisten. Dafür muss Sicherheit von Anfang an festes Element der Gestaltung des algorithmischen Systems sein (Security by Design). Das System ist vor seinem Einsatz in einer geschützten Umgebung zu testen. Die Sicherheitsvorkehrungen sind zu dokumentieren.

Konkretisierung:

Wie bei jedem Computersystem ist auch bei algorithmischen Systemen wichtig, die Sicherheit zu gewährleisten, damit sie korrekt funktionieren und unbefugte Datenzugriffe oder -manipulationen verhindert werden. Es gibt einzelne Aspekte, die sich aus dem Einsatz algorithmischer Systeme, insbesondere lernender Systeme, (neu) ergeben, wie beispielsweise die Gefahr von Manipulation im Trainingsprozess.

Bei der Gewährleistung von Sicherheit gilt es zu bedenken, dass diese nie vollständig sichergestellt werden kann, sondern es sich um einen laufenden Prozess der Überprüfung, der Behebung von Fehlern (Patchen) und der Aktualisierung von Standards handelt. Hierfür lassen sich auch viele der in anderen Bereichen der IT vorherrschenden sicherheitstechnischen Standards und Maßnahmen übernehmen.

Erfolgsfaktoren für entwickelnde Organisationen:

- Bestehende Sicherheitsrisiken bei lernenden Systemen kennen und antizipieren
- Personenbezogene Daten in algorithmischen Systemen sicher und datenschutzkonform speichern und verarbeiten
- Potenzielle Manipulationsrisiken algorithmischer Systeme erkennen
- Sicherheit komplexer werdender algorithmischer Systeme regelmäßig auf Lücken überprüfen

Orientierungsfragen für sichere Systeme:

Ermittlung des Sicherheitsstandards

- ▢ Welchen Sicherheitsrahmen braucht es für den technischen Teil und die Entwicklung des algorithmischen Systems?
- ▢ Welchen Sicherheitsrahmen braucht es für die Anwendung? Welche zusätzlichen Sicherheitsanforderungen ergeben sich aus dem soziotechnischen Kontext?

Durchführung von Sicherheitsmaßnahmen

- ▢ Wie kann von Anfang an während des Gestaltungsprozesses des algorithmischen Systems Sicherheit mitgedacht und in alle einzelnen Komponenten integriert werden (Security by Design)?
- ▢ Was sind die Anforderungen an das sichere Speichern der Daten, mit denen das algorithmische System arbeitet?
- ▢ Wie können mögliche Angriffe und Manipulationen frühzeitig erkannt werden? Wie kann eine manuelle oder automatisierte Erkennung von Angriffen und eine Notabschaltung in das algorithmische System integriert werden?

Überprüfung der Sicherheit des algorithmischen Systems

- ▢ Anhand welcher Parameter findet die Überprüfung statt?
- ▢ Wird das System vor dem Einsatz in einer geschützten Umgebung getestet, beispielsweise in einer Simulation?
- ▢ Wird das System regelmäßig im Einsatz getestet und auf Sicherheitslücken überprüft?
- ▢ Werden Sicherheitslücken aktiv gesucht und geschlossen?
- ▢ Werden Sicherheitsstandards regelmäßig überprüft und aktualisiert?

Begleitung der Sicherheitsmaßnahmen

- ▢ Wie können alle beteiligten Entwickler:innen für Sicherheitsthemen sensibilisiert werden? Wie kann ein Safety Mindset unter Entwickler:innen geschaffen werden?
- ▢ Werden Sicherheitsmaßnahmen, Gefahrensituationen und Reaktionen dokumentiert?

Weiterführende Links zu Forschung, Ansätzen und Impulsen:

Wissenschaftlicher Artikel zum Boxingverfahren für die Überprüfung lernender Systeme in einer geschützten Umgebung:
https://t1p.de/Sicherheit-Link-1

Wissenschaftlicher Artikel zur Entwicklung einer Shutdown Safety, eines Notsystems, das ein algorithmisches System bei schädlichem Input abschaltet:
https://t1p.de/Sicherheit-Link-2

5 Kennzeichnung durchführen

Der Einsatz eines algorithmischen Systems muss gekennzeichnet sein.

Beim Einsatz algorithmischer Systeme muss durch eine entsprechende Kennzeichnung für Personen, die mit ihnen interagieren, erkennbar sein, dass der Entscheidung oder Prognose ein algorithmisches System zugrunde liegt. Das gilt besonders dann, wenn das System einen Menschen in Art und Weise der Interaktion (Sprache, Aussehen etc.) imitiert.

Konkretisierung:

Eine Kennzeichnung markiert klar sichtbar diejenigen Prozesse, an denen algorithmische Systeme mitwirken. Eine empfehlenswerte Kennzeichnung ist deshalb in zwei Schritte unterteilt: Sie soll im ersten Schritt darüber informieren, ob ein algorithmisches System bei der Entscheidung oder Entscheidungsempfehlung eingesetzt wurde. Im zweiten Schritt soll die Kennzeichnung aber auch weiterführende und verständliche Informationen über die Art des Einsatzes und die Funktionsweise enthalten, die zur Nachvollziehbarkeit des Systems beitragen sollen. Die Kennzeichnung ist zudem eine Voraussetzung für Beschwerden.

Erfolgsfaktoren für entwickelnde Organisationen:

- Kennzeichnung für die Anwender:innen und Betroffenen umfassend und in einer angemessenen Form bereitstellen
- Kennzeichnung offensichtlich gestalten
- Kennzeichnung direkt mit Maßnahmen zur Nachvollziehbarkeit und Beschwerdemechanismen verknüpfen

Orientierungsfragen für eine verständliche Kennzeichnung:

Bestimmung der Anforderungen einer umfassenden Kennzeichnung

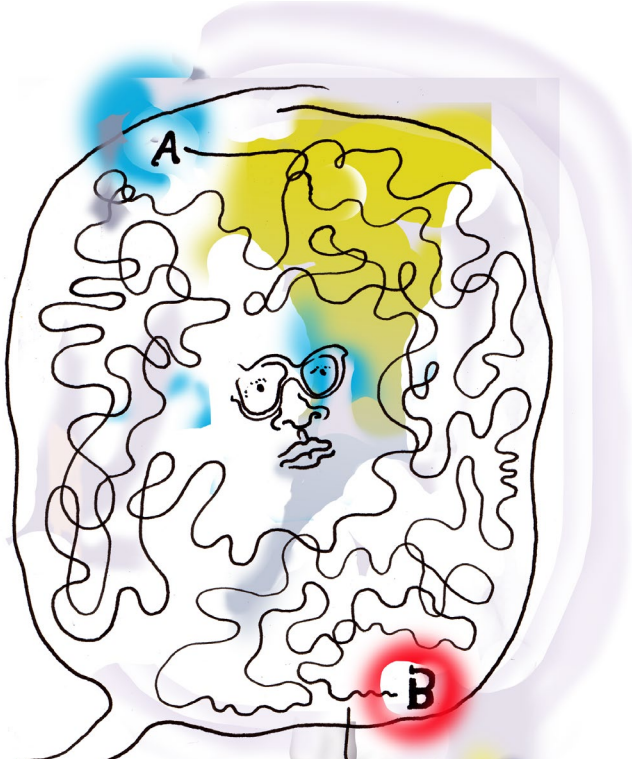
- ▢ Was sind Ansprüche der Anwender:innen und Betroffenen an eine vollständige Kennzeichnung?
- ▢ Welche Informationen sollte die Kennzeichnung beinhalten, damit sie ausführlich genug ist und alle wichtigen Aspekte abdeckt?
- ▢ Wird das algorithmische System entscheidungsunterstützend eingesetzt und welche Rolle spielt menschliche Aufsicht?
- ▢ Ist die Kennzeichnung selbst zugänglich und sind die Informationen einfach abrufbar?
- ▢ Wie wird den Anwender:innen und Betroffenen zu jedem Zeitpunkt der Nutzung des algorithmischen Systems klar gemacht, dass sie mit einem solchen System interagieren?
- ▢ Wie können sich Anwender:innen und Betroffene tiefergehend informieren?
- ▢ Welche weiteren Informationen sollten zur Sicherstellung der Nachvollziehbarkeit in die Kennzeichnung aufgenommen werden?

Bestimmung der Anforderungen an eine **allgemeinverständliche** und leicht interpretierbare Kennzeichnung

- ▢ Was sind Ansprüche der Anwender:innen und Betroffenen an eine verständliche Kennzeichnung? Wie müssen die Informationen entsprechend aufbereitet sein?
- ▢ Wurde bei der Entwicklung der Kennzeichnung auf leicht verständliche Sprache geachtet? Wurde die Nutzung von Farbgebung, Piktogrammen und/oder branchenübergreifenden Symbolen erwogen und, falls angemessen, durchgeführt?
- ▢ Wie ist die allgemeine Funktionsweise des algorithmischen Systems? Wie kann diese einfach verständlich erklärt werden?
- ▢ In welchem soziotechnischen Kontext wird das System eingesetzt? Wird der Anwendungskontext insgesamt knapp und allgemeinverständlich beschrieben?
- ▢ Welches sind die wichtigsten Attribute bzw. Faktoren, die bei der Entscheidungsfindung eine Rolle spielen? Welche Daten fließen in die Entscheidung des algorithmischen Systems ein?

Überprüfung der Kennzeichnung

- ▢ Wird bei der Kennzeichnung darauf geachtet, dass diese keine wichtigen Geschäftsgeheimnisse enthält, die eine missbräuchliche Manipulation des algorithmischen Systems ermöglichen würde?
- ▢ Wie bewerten Anwender:innen und Betroffene die Kennzeichnung im Rahmen einer Testphase und später im Einsatz? Wie kann die Kennzeichnung basierend auf deren Feedback verbessert werden?



6 Nachvollziehbarkeit sicherstellen

Die Entscheidungsfindung eines algorithmischen Systems muss stets nachvollziehbar sein.

Ein algorithmisches System muss mit seinen direkten oder mittelbaren Wirkungen und seiner Funktionsweise für Menschen leicht verständlich gemacht werden, damit diese es hinterfragen und überprüfen können. Dafür müssen Informationen über die dem System zugrunde liegenden Daten und Modelle, seine Architektur sowie die möglichen Auswirkungen veröffentlicht und in leicht verständlicher Sprache dargestellt werden. Es ist stets zu prüfen, ob sich ein Ziel ohne wesentliche Qualitätseinbußen auch mit einem algorithmischen System erreichen lässt, das weniger komplex und in seiner Funktionsweise leichter nachvollziehbar ist.

Konkretisierung:

Nachvollziehbarkeit bedeutet für Anwender:innen und Betroffene, dass sie verstehen, wie das algorithmische System funktioniert und warum es oder warum es nicht zu einem bestimmten Ergebnis gekommen ist. Nachvollziehbarkeit ist in vielen Fällen notwendig, um ein ordnungsgemäßes Funktionieren des Systems zu gewährleisten. Dies gilt insbesondere bei entscheidungsunterstützenden Systemen, wo Anwender:innen die Empfehlung, Möglichkeiten und Grenzen des Systems verstehen müssen, um es bedienen und danach angemessen entscheiden zu können. Betroffene können durch eine nachvollziehbare Anwendung Vertrauen in deren Einsatz gewinnen. Schließlich können durch Maßnahmen der Nachvollziehbarkeit mögliche negative Wirkungen erkannt und potenzielle Beschwerden durch Betroffene zielgerichtet ermöglicht werden.

Nachvollziehbarkeit kann auf unterschiedliche Art und Weise sichergestellt werden: Erstens kann ein algorithmisches System selbst nachvollziehbar sein, indem beispielsweise Modelle mit verständlichen und festen Kriterien verwendet werden. Zweitens kann das System in seiner allgemeinen Funktionsweise zusätzlich erklärt werden. Und drittens kann eine konkrete Entscheidung des Systems im Nachhinein durch nicht technische oder technische Maßnahmen nachvollziehbar gemacht oder begleitet werden. Darunter fallen beispielsweise einfache Erklärungen der wichtigen Faktoren im Einzelfall.

Erfolgsfaktoren für entwickelnde Organisationen:

- Erklärungen zielgruppenspezifisch aufbereiten
- Nachvollziehbarkeit auch besonders komplexer oder lernender Systeme in angemessenem Umfang sicherstellen
- Genügend Ressourcen für die Schaffung von Nachvollziehbarkeit von (lernenden) Systemen einplanen
- Grenzen der Sicherstellung von Nachvollziehbarkeit im Blick behalten
- Wahrung der Geschäftsgeheimnisse mit Ansprüchen an Transparenz in Einklang bringen

Orientierungsfragen für nachvollziehbare Systeme und deren transparenten Einsatz:

Ermittlung der Erfordernisse an Nachvollziehbarkeit

- ▢ Was sind, mit dem technischen Entwicklungsstand im Blick, messbare Ziele für die Sicherstellung der Nachvollziehbarkeit für unterschiedliche Zielgruppen?
- ▢ Kann das algorithmische System seine Ziele auch mit weniger komplexen und einfacher verständlichen Modellen erreichen? Welches ist die nachvollziehbarste Option?
- ▢ Wurden die Erfordernisse und die Ziele im Austausch mit Stakeholdern ermittelt und reflektiert?
- ▢ Wurden externe Expert:innen einbezogen, um die für diesen Anwendungskontext passende Methode zur Sicherstellung der Nachvollziehbarkeit auszuwählen?
- ▢ Wie viel Zeit, Vorwissen und Bereitschaft bringen Anwender:innen und Betroffene mit, um sich mit dem algorithmischen System auseinanderzusetzen?
- ▢ Von welchen anderen Organisationen und Ansätzen kann ich mich inspirieren lassen?

Erklärung der **allgemeinen Funktionsweise**

- ▢ Welche wichtigen Kriterien spielen bei der Entscheidung des algorithmischen Systems eine Rolle? Welches Gewicht haben diese Kriterien jeweils?
- ▢ Welches sind die Grenzen des algorithmischen Systems? Was kann es leisten, was nicht?
- ▢ Was sind die Ziele und der Einsatzkontext des algorithmischen Systems?

Erklärung der **konkreten Entscheidung im Einzelfall**

- ▢ Mithilfe welcher technischen Maßnahmen können im Nachgang der Entscheidung die relevanten Faktoren ermittelt werden? Was ist der aktuelle wissenschaftliche Stand dazu?
- ▢ Wie könnten Externe, wie z. B. wissenschaftliche Einrichtungen, dabei helfen, technische Maßnahmen zur Nachvollziehbarkeit zu treffen?
- ▢ Wie kann den Anwender:innen und Betroffenen kommuniziert werden, welche Faktoren für die sie persönlich betreffende Entscheidung relevant waren?
- ▢ Wie können diese Informationen durch Texte und Graphiken so aufbereitet werden, dass sie leicht verständlich und interpretierbar sind?
- ▢ Wie können diese Informationen den Anwender:innen und Betroffenen niedrigschwellig zugänglich gemacht werden, z.B. als Teil der Kennzeichnung?

Laufende Überprüfung und **Anpassung**

- ▢ Wird die Erreichung der messbaren Ziele regelmäßig überprüft?
- ▢ Werden laufend Tests durchgeführt, um die Verständlichkeit der Erklärungen und die Fähigkeit von Externen, das System zu hinterfragen, sicherzustellen?
- ▢ Welche unterschiedliche Zielgruppen werden dabei befragt?
- ▢ Welche Anpassungen an den Maßnahmen müssen durchgeführt werden?

Weiterführende Links zu Forschung, Ansätzen und Impulsen:

Blogartikel zum Einsatz von LIME zur Herstellung von Nachvollziehbarkeit:
https://t1p.de/Nach-Link-1

Webseite, die Ressourcen zu LRP sammelt, einem Ansatz, der Nahvollziehbarkeit durch Umkehrung eines neuronalen Netzes herstellt:
https://t1p.de/Nach-Link-2 https://t1p.de/Nach-Link-3

Wissenschaftlicher Artikel zur Methode der Counterfactual Explanations:
https://t1p.de/Nach-Link-4

Studie zur Nachvollziehbarkeit algorithmischer Systeme für Betroffene:
https://t1p.de/Nach-Link-5

7 Beherrschbarkeit absichern

Ein algorithmisches System muss während seines gesamten Einsatzes gestaltbar sein und bleiben.

Damit ein algorithmisches System gestaltbar bleibt, müssen alle Personen, die an seiner Entwicklung und seinem Einsatz beteiligt sind, gemeinsam stets die Kontrolle über das System behalten. Dabei muss sichergestellt werden, dass der Gesamtüberblick über das System stets gewahrt bleibt, auch wenn Aufgaben zwischen unterschiedlichen Personen und Arbeitsbereichen verteilt sind. Die Arbeitsweise eines Systems darf niemals so komplex und undurchschaubar werden, dass es von Menschen nicht mehr beherrschbar ist oder nicht mehr geändert werden kann. Das gilt insbesondere für selbstlernende Systeme. Kann diese Beherrschbarkeit nicht sichergestellt werden, so ist auf den Einsatz eines algorithmischen Systems zu verzichten.

Konkretisierung:

Beherrschbarkeit im Sinne der Algo.Rules umfasst zwei Dimensionen. Erstens Beherrschbarkeit als Justierbarkeit: Das System muss bzgl. der Zielvorgaben veränderbar sein. Können Komponenten nicht oder nur schwer veränderbar sein, kann es auch ausreichen, sie austauschen zu können. Zweitens Beherrschbarkeit als Abschaltbarkeit: Ein System muss technisch stets abschaltbar bleiben. Um das zu gewährleisten, braucht es auch organisatorische Maßnahmen.

Beherrschbarkeit ist erforderlich, um der Verantwortung für das algorithmische System nachkommen zu können. Beherrschbarkeit bedeutet deshalb auch, dass Fehler im algorithmischen System, die einmal durch Überprüfungsmechanismen entdeckt oder durch Beschwerden gemeldet wurden, behoben werden. Sie ist wichtig, damit algorithmische Systeme nicht nur menschengemacht sind, sondern auch menschengestaltbar bleiben.

Erfolgsfaktoren für entwickelnde Organisationen:

- Personen, die die Beherrschbarkeit algorithmischer Systeme sicherstellen sollen, mit entsprechender Fachkenntnis und Ressourcen ausstatten
- Auch in besonders großen Teams Verantwortung für Beherrschbarkeit klar definieren
- Wissen zu Beherrschbarkeit umfassend dokumentieren
- Abschaltbarkeit vorausschauend vorbereiten

Orientierungsfragen für beherrschbare algorithmische Systeme:

Bestimmung der Anforderungen an die Beherrschbarkeit

- ▢ Welche Anforderungen haben Betroffene, Anwender:innen, Entwickler:innen und Führungskräfte jeweils an die Justierbarkeit des algorithmischen Systems?
- ▢ In welchen Fällen sollte es – aus der Sicht unterschiedlicher Personen(gruppen) - zu einer Abschaltung des algorithmischen Systems kommen? Inwiefern sind bestimmte Personen(gruppen) von dem Funktionieren des Systems abhängig?

Sicherstellung der **Justierbarkeit** des algorithmischen Systems

- ▢ Wie kann bereits im Planungsprozess des algorithmischen Systems auf Justierbarkeit geachtet werden? Welche Modelle wären am ehesten justierbar?
- ▢ Ist das algorithmische System, soweit möglich, anpassbar und modularisiert aufgebaut?
- ▢ Wer ist in welcher Weise für Anpassungen des algorithmischen Systems zuständig?
- ▢ Wird das Wissen über die Justierbarkeit des algorithmischen Systems laufend weitergegeben bzw. zentral zugänglich gemacht?

Dokumentation des Entwicklungsprozesses

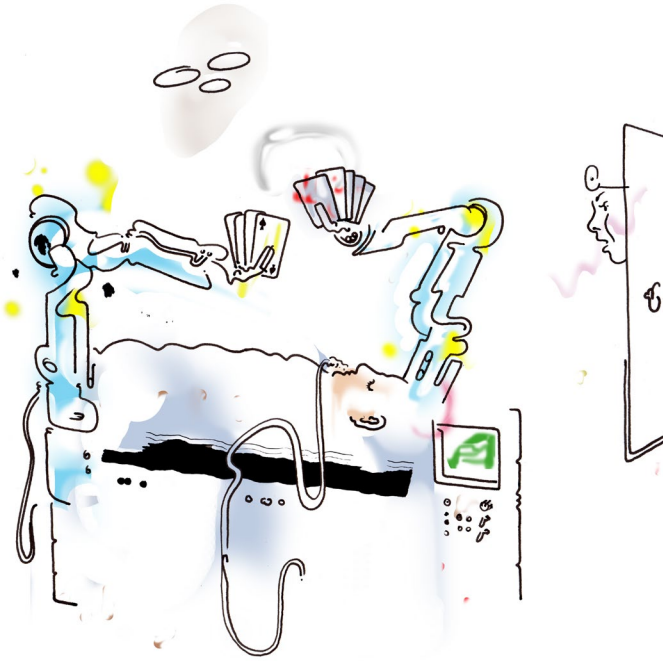
- ▢ Gibt es einen Überblick aller Module bzw. technischen Elemente des Systems und durchgeführten Tests, auch über Produktversionen und beteiligte Teams hinweg?
- ▢ Was sind die Abhängigkeiten und Interaktionen zwischen einzelnen Elementen?
- ▢ Wie werden Änderungen des algorithmischen Systems erfasst? Welche automatisierten und händischen Prozesse können dafür aufgesetzt werden?
- ▢ Wie kann sichergestellt werden, dass einzelne Prozessschritte nachvollziehbar bleiben?
- ▢ Wie werden die ursprünglich identifizierten Ziele des Systems implementiert?

Sicherstellung der **technischen Abschaltbarkeit** des algorithmischen Systems

- ▢ Anhand welcher Indikatoren kann ich erkennen, dass das algorithmische System abgeschaltet werden muss?
- ▢ Welche Auswirkungen hätte die Abschaltung für verschiedene Stakeholder? Welche Risiken wären mit einer Abschaltung verbunden? Wie können diese Risiken minimiert werden?
- ▢ Wer würde über die Abschaltung eines Systems entscheiden?
- ▢ Welches wären die technischen und organisatorischen Schritte zur Abschaltung des Systems?
- ▢ Welche alternativen Systeme müssten aufgesetzt oder vorbereitet werden, damit diese im Falle der Abschaltung die Funktion des algorithmischen Systems ersetzen könnten?
- ▢ Wie würden Anwender:innen und Betroffene über die Abschaltung informiert werden?

Weiterführende Links zu Forschung, Ansätzen und Impulsen:

Wissenschaftlicher Artikel zu Justierungsmethoden lernender Systeme:
https://t1p.de/Beherrschbarkeit-Link-1



8 Wirkung überprüfen

Die Auswirkungen eines algorithmischen Systems müssen regelmäßig überprüft werden.

Ein algorithmisches System muss einer aktiven Kontrolle unterliegen, ob die beabsichtigten Ziele tatsächlich verfolgt werden und der Einsatz des Systems bestehendes Recht nicht verletzt. Externe Prüfstellen sollten unter Wahrung legitimer Geschäftsgeheimnisse durch entsprechende technische Vorkehrungen in die Lage versetzt werden, ein algorithmisches System tatsächlich und umfassend unabhängig prüfen zu können. Wird eine negative Wirkung festgestellt, muss die Fehlerursache ermittelt und das algorithmische System entsprechend angepasst werden.

Konkretisierung:

Die Wirkungsüberprüfung analysiert, welche tatsächlichen Auswirkungen das algorithmische System in der konkreten Anwendung hat. Die laufende Überprüfung soll sicherstellen, dass Fehlerursachen erkannt und behoben werden können. Das können sowohl Anpassungen des Codes oder anderer technischer Komponenten des algorithmischen Systems als auch Anpassungen des soziotechnischen Kontexts sein.

Die laufende und kontinuierliche Überprüfung sollte stets kontextspezifisch und in Hinblick auf konkrete Auswirkungen auf Stakeholder und Betroffene sowie auf die Gesellschaft vorgenommen werden. Eine regelmäßige Kontrolle durch externe Prüfstellen soll sicherstellen, dass die Ergebnisse unabhängig ermittelt werden. Auf technischer Ebene greifen Maßnahmen zur Wirkungsüberprüfung mit Maßnahmen zur Schaffung von Nachvollziehbarkeit eng ineinander. Dabei kann eine Wirkungsüberprüfung auch stattfinden, ohne dass der Code oder andere Geschäftsgeheimnisse offenbar gemacht werden müssen.

Erfolgsfaktoren für entwickelnde Organisationen:

- Wirkungen messbar machen und dafür ein oder mehrere Qualitätsmaße festlegen
- Prüfkriterien gemeinschaftlich entwickeln
- Neben quantitativen, technischen Wirkungsüberprüfungsverfahren auch qualitative und sozialwissenschaftliche Methoden anwenden
- Wirkungsüberprüfung des Gesamtsystems auch bei komplexen Systemen ganzheitlich durchführen
- Wirkung regelmäßig oder laufend überprüfen, insbesondere bei lernenden Systemen und sich verändernden Umgebungen
- Wirkungsüberprüfung langfristig denken und auch nach der Abnahme der Software durchführen
- Wirkung unter Wahrung von Geschäftsgeheimnissen extern prüfen lassen
- Mögliche diskriminierende Wirkungen antizipieren und vorbeugen

Orientierungsfragen für eine umfassende Wirkungsüberprüfung:

Überprüfung der Zielvorgaben des algorithmischen Systems

- ▢ Was sind die aus den Zielvorgaben resultierenden Anforderungen (Requirements) für Testverfahren?
- ▢ Wann werden die Zielvorgaben erfüllt? Was sind die entsprechenden Qualitätsmaße und Mindestanforderungen?
- ▢ Inwiefern werden die in Algo.Rule 3 identifizierten Ziele und Wirkungen und die bei der Wirkungsrelevanz ermittelten positiven wie negativen Auswirkungen erreicht?

Interne Überprüfung der Wirkung des algorithmischen Systems

- ▢ Wer ist für die Kontrolle des algorithmischen Systems intern zuständig?
- ▢ Wie wird der Programmcode überprüft?
- ▢ Werden bei lernenden Systemen nicht nur das ausführende algorithmische System, sondern auch der Lernalgorithmus und seine Hyperparameter während des Trainingsprozesses überprüft?
- ▢ Wie werden die Inputdaten, Outputdaten und die Interaktion der Anwender:innen mit dem System überprüft?
- ▢ Können synthetische Daten und Simulation von Anwender:innen dabei helfen, die Wirkung abzuschätzen?
- ▢ Wie kann in der konkreten, individuellen Anwendung die Wirkungen auf Anwender:innen oder Betroffene sowie die Gesellschaft überprüft werden?
- ▢ Werden die Testverfahren und Überprüfungsprozesse sowie mögliche Eingriffe in das algorithmische System dokumentiert?

Externe Überprüfung der Wirkung des algorithmischen Systems

- ▢ Welches Überprüfungsintervall ist der Dynamik des Systems und des Anwendungsumfelds angemessen?
- ▢ Welche externe Stelle könnte eine solche Überprüfung durchführen?
- ▢ Welche Informationen und Zugänge brauchen diese externen Stellen, um eigenständig Überprüfungen vorzunehmen? Wie können Zugänge und Audits aussehen, die gleichzeitig legitime Geschäftsgeheimnisse wahren?
- ▢ Werden externe Stakeholder zu den Auswirkungen befragt?

Weiterführende Links zu Forschung, Ansätzen und Impulsen:

Überblick über verschiedene Auditmethoden algorithmischer Systeme:
https://t1p.de/Wirkung-Link-1

Forschungsprojekt zu einem AI Ethics Inspection Process, um u. a. die Wirkungen lernender Systeme zu überprüfen:
https://t1p.de/Wirkung-Link-2

Forschungsprojekt zu prädiktiven Datenverarbeitungsverfahren, das systematische Schwachstellen in Entscheidungssystemen aufdecken soll:
https://t1p.de/Wirkung-Link-3

Thesenpapier, das u. a. aus geltendem Recht den Bedarf auf externe Einsichtnahme und Kontrolle algorithmischer Systeme rechtfertigt:
https://t1p.de/Wirkung-Link-4

Forschungspapier, das Auditmechanismen für algorithmische Systeme beschreibt:
https://t1p.de/Wirkung-Link-5

9 Beschwerden ermöglichen

Fragwürdige oder die Rechte einer betroffenen Person beeinträchtigende Entscheidungen eines algorithmischen Systems müssen erklärt und gemeldet werden können.

Die ein algorithmisches System einsetzende Stelle muss leicht zugängliche Wege zur Kontaktaufnahme zur Verfügung stellen. Betroffene Personen müssen erstens eine qualifizierte und detaillierte Auskunft zur konkreten Entscheidung und der dahinter liegenden Abwägung einfordern können. Diese Möglichkeit sollte auch in ihrem Namen handelnden oder eigene legitime Interessen geltend machenden Organisationen eingeräumt sein. Zweitens muss eine einfache, niedrigschwellige und wirksame Beschwerdemöglichkeit zur Verfügung stehen. Beschwerden und eingeleitete Maßnahmen sind zu dokumentieren.

Konkretisierung:

Beschwerden sind Hinweise auf mögliche Fehler oder negative Auswirkungen eines Systems im Einsatz. Diese können dabei sowohl von Externen kommen als auch intern eingereicht werden. Ein effektiver Beschwerdemechanismus ermöglicht, dass jede Beschwerde überprüft und beantwortet wird, sodass die Systeme laufend verbessert werden können. Auch im Planungszeitraum nicht antizipierte Wirkungen können im realen Einsatz so erfasst und korrigiert werden. Das trägt zum Vertrauen der Anwender:innen und Betroffenen in das algorithmische System bei. Dafür ist es notwendig, dass die Beschwerde einfach, niedrigschwellig und ohne hohe Voraussetzungen eingereicht werden kann.

Im Falle einer legitimen Beschwerde müssen Prozesse der Fehleranalyse und Korrektur starten. Mögliche negative Folgen des Fehlers müssen nicht nur ausgeglichen, sondern ebenso in Zukunft vermieden werden. Da in einigen Fällen benachteiligte Gruppen einem Diskriminierungsrisiko ausgesetzt sein können, sollte eine Beschwerde auch als Gruppe eingereicht werden können. Legitime Interessen geltend machende Organisationen, wie Interessenvertretungen, Verbraucher:innenschutzverbände oder Antidiskriminierungsstellen, sollten deshalb ebenfalls Zugang zum Beschwerdemechanismus haben.

Erfolgsfaktoren für entwickelnde Organisationen:

- Beschwerdemöglichkeit für Betroffene niedrigschwellig gestalten
- Beschwerdeprozess so gestalten, dass Gestalter:innen algorithmischer Systeme anhand der Beschwerden Fehlerquellen identifizieren können
- Relevante Beschwerden identifizieren und herausfiltern
- Negative Folgen für besonders marginalisierte Personengruppen im Blick behalten
- Beschwerdeprozess institutionalisieren, beispielsweise durch die Einrichtung einer Ombudsperson

Orientierungsfragen für effektive Beschwerdemechanismen:

Technische Vorbereitung der Beschwerdeverarbeitung

- ▢ Ist es sinnvoll, laufend bzw. pro Session die Vorgänge des algorithmischen Systems dokumentiert und als Logdaten zwischenzuspeichern?
- ▢ Welche Informationen sollten bereitgehalten werden, um diese im Fall von Beschwerden an Betroffene kommunizieren zu können?

Sicherstellung der **Zugänglichkeit** des Beschwerdemechanismus

- ▢ Ist der Beschwerdemechanismus intuitiv auffindbar und bedienbar sowie barrierearm?
- ▢ Wie kann der Beschwerdemechanismus sinnvoll an andere Maßnahmen gekoppelt werden, wie jene zur Kennzeichnung und Nachvollziehbarkeit, um Informationen für Anwender:innen und Betroffene zu bündeln?
- ▢ Gibt es eine einfache Möglichkeit, Beschwerden inkl. Logdaten zu versenden?
- ▢ Welche weiteren Informationen können Betroffene zur Verfügung stellen?
- ▢ Wurde der Beschwerdemechanismus gemeinsam mit den Stakeholdern entwickelt?

Interne Verarbeitung der Beschwerden

- ▢ Wer ist für Beschwerden und ihre Bearbeitung zuständig?
- ▢ Braucht es möglicherweise eine gesonderte, unabhängige Stelle für interne und/oder externe Beschwerden?
- ▢ Welchen Prozessablauf gibt es für die Beschwerdeverarbeitung? Wie können sich Betroffene möglicherweise direkt äußern?
- ▢ Wie und anhand welcher Kriterien wird entschieden, welche Beschwerde weiterverfolgt wird?
- ▢ Wie werden Beschwerden beantwortet? Welche Informationen werden Betroffenen im Fall eine Beschwerde zur Verfügung gestellt? Wie werden auf Grundlage der Beschwerde getroffene Maßnahmen den Betroffenen kommuniziert?
- ▢ Wie wird (intern oder extern) überprüft, ob und wie Beschwerden verarbeitet werden?

Behebung der durch Beschwerden identifizierten Probleme

- ▢ Wie wird mit dem Einzelfall umgegangen? Wird der Fall durch einen Menschen erneut überprüft?
- ▢ Inwiefern werden mögliche Schäden erstattet, korrigiert oder behoben?
- ▢ Werden beispielsweise eine Fehleranalyse, Fehlerquellenidentifikation und eine Anpassung des algorithmischen Systems durchgeführt, um solche Fehler in Zukunft zu vermeiden?
- ▢ Welche Ähnlichkeiten gibt es zwischen unterschiedlichen Beschwerden? Wie kann eine systematische Prüfung der Auswirkungen durchgeführt werden, um unabhängig von Einzelfällen gruppenbezogene Auswirkungen zu erfassen?

Impressum und Kontakt

Dieser Leitfaden wurde erstellt im Rahmen des Projekts Algo.Rules, das die Bertelsmann Stiftung gemeinsam mit dem iRights.Lab umsetzt.

Stand: Juni 2020

Bertelsmann Stiftung
Carl-Bertelsmann-Straße 256
33311 Gütersloh
www.bertelsmann-stiftung.de
Verantwortlich: Carla Hustedt und Lajla Fetic

Umgesetzt mit:
Think Tank iRights.Lab
Schützenstraße 8
10117 Berlin
www.irights-lab.de
kontakt@irights-lab.de
Verantwortlich: Philipp Otto

Autor:innen: Michael Puntschuh, Lajla Fetic
Redaktion: Wiebke Glässer, Jaana Müller-Brehm, Ramak Molavi, Gina Schad
Lektorat: Julia Schrader, Rudolf Jan Gajdacz / team 4media&event

Die Algo.Rules werden in einem offenen, partizipativen und interdisziplinären Prozess mit bislang über 400 Beteiligten erarbeitet. Die Inhalte dieses Arbeitspapiers wurden insbesondere durch einen Workshop und Konsultationen mit Entwickler:innen und weiteren Expert:innen erarbeitet. Wir danken allen Beteiligten für ihren Input.

Kontakt

Falls Sie Fragen haben oder sich näher zu dem Thema austauschen wollen, können Sie sich via algorules@irights-lab.de bei uns melden. Wir haben neben den Algo.Rules ein Impulspapier für Führungskräfte veröffentlicht und arbeiten aktuell an Praxishilfen für den öffentlichen Sektor. Weitere Informationen zum Projekt finden Sie auf der Webseite www.algorules.org.

Lizenz

DOI 10.11586/2020029