2 Umgang mit Daten in Datenökosystemen

Dieser erste Baustein meiner Argumentation schafft die Grundlage für das Vorhaben zu zeigen, wie der Umgang mit digitalen Daten in Datenökosystemen ethisch gestaltet werden kann. Dafür wird zuerst einmal definiert was Daten sind und wie diese mit Hilfe von Algorithmen verarbeitet werden können. Dann definiere ich Datenökosysteme über den Datenfluss zwischen vier Akteuren – Datenquellen, Datengenerierenden, Datenverarbeitenden und Datennutzenden – mit verschiedenen Rollen und stelle ein eigenes Modell von der Struktur eines Datenökosystems auf. Dabei zeigt sich, dass das Sammeln und Verarbeiten von Daten sowohl Vorteile hat, als auch Risiken birgt. Dieser für meine Argumentation zentrale Begriff des Risikos wird definiert und ich zeige, dass im Datenökosystem nicht alle Akteure über das gleiche Wissen verfügen bezüglich der Datenverarbeitung im Datenökosystem und dass ungleiche Kontrollmöglichkeiten vorherrschen, weshalb auch die Risiken ungleich verteilt sind.

2.1 Datenverarbeitung

Um sich mit der ethischen Bewertung und Gestaltung von Datenverarbeitung in Datenökosystemen auseinandersetzen zu können, muss zuerst einmal geklärt werden, was genau mit Daten gemeint ist. Die Technologieforscher Syed Iftikhar Hussain Shah, Vasilis Peristeras und Ioannis Magnisalis definieren Daten folgendermaßen: „Data are facts and figures about an object (…).“[88] Die Daten, von denen in dieser Arbeit die Rede ist, sind digital gespeichert und sie liegen ursprünglich in roher Form vor, zum Beispiel als Dokumente, Transkripte oder audio-visuelle Aufnahmen.[89] Die Daten können dabei

88 Shah et al. 2020:102
89 Poikola et al. 2011:13

über alles Mögliche eine Aussage machen, zum Beispiel über die Umwelt, Gegenstände oder Personen. In dieser Arbeit sind vor allem die Daten von Interesse, die eine Aussage über Personen machen – ich nenne sie personenbezogene Daten. Welche Daten personenbezogen sind, ist umstritten, aber in dieser Arbeit wird in Übereinstimmung mit der Datenschutzgrundverordnung (DSGVO) angenommen, dass alle Daten, die eine Aussage über eine bestimmte Person machen, personenbezogene Daten sind.[90] Personenbezogene Daten offenbaren also direkt oder indirekt die Identität von bestimmten Personen – hier kann es sich um den Namen, die Anschrift oder auch um Fotos der Person handeln.[91]

Personenbezogene Daten können dann als private Daten bezeichnet werden, wenn die Personen, über welche die Daten eine Aussage machen, diese Daten nicht der Öffentlichkeit preisgeben möchten. Eine Teilmenge der privaten Daten wird als sensible Daten bezeichnet, wenn die betroffenen Personen sie nicht öffentlich machen wollen, weil sie bei einer Offenbarung mit negativen Auswirkungen rechnen. Dies betrifft Daten, die Auskünfte enthalten über ethnische Herkunft, politische Meinungen, religiöse Überzeugungen, Gesundheit oder sexuelle Orientierung einer bestimmten Person.[92] Dabei gelten solche Daten als direkte Identifikatoren, die eine Person unmittelbar erkennbar machen, wie zum Beispiel Name, Anschrift oder Fotos. Als indirekte Identifikatoren werden solche Daten bezeichnet, die nicht direkt auf die Identität einer Person hinweisen, aber durch die Kombination mit anderen Daten auf eine Person schließen lassen. So können zum Beispiel auch die Postleitzahl und das Alter indirekte Identifikatoren sein, wenn zum Beispiel in einem Ort nur eine 90-jährige Frau wohnt.[93]

Um große Mengen von Daten verstehen oder interpretieren zu können, müssen sie analysiert werden. Damit rohe Daten zur Verarbeitung bereit sind, werden sie homogenisiert und in ein maschinenlesbares Format gebracht.[94] Eine solche Analyse kann erleichtert werden, durch den Einsatz von Algorithmen, die bestimmte Muster

90 Art. 4, Abs. 1, DSGVO
91 Zainab, Kechadi 2019:2
92 Art. 9, Abs. 1 DSGVO; Zainab, Kechadi 2019:2f.
93 Zainab, Kechadi 2019:3
94 Poikola et al. 2011:29, 32

in den Daten erkennen. So kann durch eine algorithmische Analyse aus Daten Wissen gewonnen werden.[95] Denn die Datenverarbeitung erzeugt neue Informationen, die nicht mehr als Rohdaten bezeichnet werden können.[96] Wenn sehr große Datenmengen analysiert werden sollen, benötigt man neue Methoden der Datenverarbeitung – bei solch großen Mengen von Daten spricht man von Big Data[97], und der Prozess der Wissensgewinnung, -darstellung und -anwendung nennt sich Data Mining.[98] Durch das Erkennen von Mustern in diesen vielen Datensätzen wird es möglich, Aussagen über ihren Inhalt zu treffen, die wirtschaftlich, sozial, technisch oder rechtlich relevant sein können. Der Glaube, dass große Mengen an Datensätzen zu Einsichten führen können, die zuvor unmöglich waren, ist laut Medienwissenschaftlerin und Sozialforscherin Danah Boyd sowie KI-Forscherin Kate Crawford jedoch ein Mythos.[99] Die bereits erwähnten Technologiephilosoph:innen Solon Barocas und Helen Nissenbaum aber halten dagegen, dass Erkenntnisse durch eine Analyse großer Datenmengen unerwartet sein können und dass darin der Wert der Daten liegt.[100]

Diese Analyse großer Datenmengen wird durch Algorithmen ermöglicht. Die Informatiker Thomas H. Cormen, Charles E. Leiserson, Ronald Rivest und Clifford Stein definieren Algorithmen wie folgt: „Somit ist ein Algorithmus eine Folge von Rechenschritten, die die Eingabe in die Ausgabe umwandeln."[101] Es handelt sich bei einem Algorithmus um ein Hilfsmittel zur Lösung von Rechenproblemen. Zum Beispiel können Zahlen in eine andere Reihenfolge sortiert werden. In der Eingabe sind die Zahlen unsortiert, in der Ausgabe sollen sie aber in aufsteigender Reihenfolge erscheinen. Der Algorithmus enthält die Angaben, die notwendig sind, um die Zahlen in aufsteigender Reihenfolge zu sortieren. Dafür muss definiert sein, was mit „in aufsteigender Reihenfolge" gemeint ist. Cormen et al. fügen hinzu: „Ein Algorithmus wird als korrekt bezeichnet, wenn

95 Immonen et al. 2014:89
96 Poikola et al. 2011:74
97 Shah et al. 2020:102; Altintac 2022:5; Corsten, Roth 2022:4; Donges 2022:214
98 Ertel 2021:206
99 Boyd, Crawford 2012:663
100 Barocas, Nissenbaum 2014:60
101 Cormen et al. 2013:5

er für jede Eingabeinstanz mit der korrekten Ausgabe stoppt."[102] Ein korrekter Algorithmus löst das gegebene Rechenproblem und würde in diesem Fall die eingegebenen Zahlen in der definierten Weise in aufsteigender Reihenfolge sortieren. Ein Algorithmus ist also eine präzise und eindeutige Handlungsanweisung, welche dazu führt, dass der ihn nutzende Mensch aus der Eingabe die erwartete Ausgabe erhält.

So kann zum Beispiel mit Hilfe eines Algorithmus schnell ausgerechnet werden, ob Antragstellende ein Anrecht auf Wohngeld haben und in welcher Höhe ihnen dieses Geld nach dem Wohngeldgesetz (WoGG) zusteht. Dafür müssen die Bedingungen des Wohngeldgesetzes in den Algorithmus einprogrammiert werden. Es muss also konkret vorgegeben werden, inwiefern sich das Wohngeld nach der Anzahl der zu berücksichtigenden Haushaltsmitglieder, der zu berücksichtigenden Miete oder Belastung und dem Gesamteinkommen richtet und unter welchen Bedingungen man keinen Anspruch auf Wohngeld hat.[103] Dann kann dieser Algorithmus aus den Daten von Antragstellenden berechnen, ob ein Anspruch besteht und in welcher Höhe. In dieser Arbeit nenne ich solche Algorithmen Wenn-Dann-Algorithmen. Denn damit wird eine Bedingung klar vorgegeben, zum Beispiel *wenn* „das Wohngeld weniger als 10 Euro monatlich betragen würde"[104], *dann* besteht kein Wohngeldanspruch. Solche Wenn-Dann-Algorithmen können Rechenprobleme schneller lösen als Menschen und diese somit zum Beispiel bei der Bearbeitung von Wohngeldanträgen entlasten.

Die „Digitization" hat einen solchen Einsatz von Algorithmen zur Problemlösung ermöglicht. Während es bei der Digitalisierung zuerst nur um die Automatisierung von Prozessen ging, geht es mittlerweile um die Autonomisierung von Prozessen.[105] Das heißt, dass inzwischen technische Systeme selbstständig lernen und Lösungen anbieten sollen. Diese Systeme werden als Künstliche Intelligenz (KI) bezeichnet. Die Bezeichnung „Künstliche Intelligenz" stammt aus einem Antrag für ein Forschungsprojekt von den Informatikern und Mathematikern John McCarthy, Marvin Minsky, Nathaniel Roches-

102 Cormen et al. 2013:6
103 §4 – 21 WoGG
104 §21 WoGG
105 Klenk et al. 2020:5f.

ter und Claude Elwood Shannon im Jahr 1955. Damals schrieben sie: „The study is to proceed on the basis of the conjecture that every aspect of learning or any other feature of intelligence can in principle be so precisely described that a machine can be made to simulate it."[106] Das Ziel war, menschliche Intelligenz von einer Maschine simulieren zu lassen. Diese Maschine sollte Probleme lösen, Konzepte formulieren, Sprache verwenden und sich selbst verbessern können. Einer der Forscher, Marvin Minsky, definiert KI so: „(…) *artificial intelligence*, the science of making machine do things that would require intelligence if done by men."[107] Das Ziel der Wissenschaftler war es, Maschinen dazu zu bringen, Dinge zu tun, die Intelligenz erfordern würden, wenn sie von Menschen ausgeführt würden.

Da die Verwendung des Intelligenzbegriffs uns in dessen kontroverse Definitionen hineinzieht, bietet es sich an, alternative Definitionen von KI zu betrachten. So zum Beispiel die Definition der Informatikerin Elaine A. Rich: „Artificial Intelligence is the study of how to make computers do things at which, at the moment, people are better."[108] Die Definition umgeht die inhaltliche Aufschlüsselung des Intelligenzbegriffs, schreibt Intelligenz aber grundsätzlich dem Menschen zu, und gleichzeitig hält er diesbezüglich die aktuelle Überlegenheit des Menschen gegenüber dem Computer fest. Konkreter formulieren dies die Informatiker Uwe Lämmel und Jürgen Cleve: Künstliche Intelligenz ist ein „Teilgebiet der Informatik, welches versucht, menschliche Vorgehensweisen der Problemlösung auf Computern nachzubilden, um auf diesem Wege neue oder effizientere Aufgabenlösungen zu erreichen."[109] Intelligenz wird hier gleichgesetzt mit menschlicher Problemlösung. Die Definition geht allerdings darüber hinaus und erwartet von der Künstlichen Intelligenz neue und effizientere Vorgehensweisen.

Daraufhin stellt sich die Frage, wozu die Künstliche Intelligenz fähig ist. Aktuell gibt es nur sogenannte schwache KI (Weak AI), das heißt Programme, die Fähigkeiten der Intelligenz demonstrieren, ohne selbst „intelligent" zu sein. Viel diskutiert wird auch die Entwicklung von starker KI (Strong AI), welche ein Programm sein

106 McCarthy et al. 1955:2
107 Minsky 1968:v (Preface)
108 Rich 1983; Vgl. Ertel 2021:2
109 Lämmel, Cleve 2008:14

soll, das ein Bewusstsein hat, was aktuell nicht existiert.[110] Die Frage, ob Computerprogramme irgendwann Bewusstsein erlangen können oder sollten, fällt unter die von Bolte und van Wynsberghe sogenannte erste Welle der KI-Ethik und soll hier nicht weiter behandelt werden.[111] In dieser Arbeit ist mit KI immer schwache KI gemeint, und ich gehe davon aus, dass Computerprogramme nur einem Regelsatz folgen, um Symbole abgleichen und Input in Output verwandeln zu können.

Wer den Regelsatz bestimmt, unterscheidet sich bei verschiedenen Typen von Algorithmen. Während bei Wenn-Dann-Algorithmen der Regelsatz klar vom Menschen vorgegeben wird, wird bei KI-Systemen[112] nur noch das Ziel des Outputs vorgegeben. Die dabei der Maschine vorgegebenen Werte oder Ziele sollen mit denen der Menschen übereinstimmen – da dies nicht immer einfach ist, nennt sich der Versuch, diese Übereinstimmung herzustellen, das Problem der Werteanpassung (value alignment).[113] Der Regelsatz für die Symbolverarbeitung kann also nicht nur vorgegeben werden, sondern er kann auch durch KI-Systeme erschlossen werden – diese Methode nennt sich Machine Learning (ML; Maschinelles Lernen). Hierbei handelt es sich um ein Teilgebiet der Künstlichen Intelligenz, welches statistische Methoden verwendet, um Maschinen in die Lage zu versetzen, sich durch Erfahrung zu verbessern.[114] Beim maschinellen Lernen liest ein Computer einige Daten, erstellt ein Modell auf der Grundlage dieser Daten und verwendet das Modell sowohl als These über die Welt als auch als Software, die Probleme lösen kann. Die Informatiker Stuart J. Russel und Peter Norvig definieren dieses Lernen so: „An agent is learning if it improves its performance after making observations about the world."[115] Dem liegt das Prinzip der

110 Searle 1980:417; Boddington 2023:3f.
111 Bolte, van Wynsberghe 2024:1735f
112 Hierbei handelt es sich nicht um ein soziotechnisches System, wie bei Datenökosystemen, sondern um ein rein technisches System. Das heißt, hier stehen verschiedene technische Objekte in bestimmten Beziehungen zueinander (Shapiro 1997:73).
113 Russel, Norvig 2021:22
114 Mitchell 1997:XV; Ertel 2021:201
115 Russel, Norvig 2021:669; Hervorhebung entfernt

Induktion zugrunde, da aus wiederkehrenden Mustern in Datensätzen allgemeine Regeln abgeleitet werden.[116]

Es gibt verschiedene Ansätze des Maschinellen Lernens, man unterscheidet supervised learning, unsupervised learning und reinforcement learning. Beim supervised learning (Überwachtes Lernen) beobachtet das KI-System Eingabe-Ausgabe-Paare und erlernt die Funktion, welche Eingabe und Ausgabe verbindet. Dabei werden die Eingabe-Ausgabe-Paare von den Programmierenden vorgegeben, zum Beispiel Fotos (Eingabe), die beschriftet werden mit den Dingen, die auf ihnen abgebildet sind, wie „Katze" oder „Hund" (Ausgabe). Solche Ausgaben werden als „Label" bezeichnet und ermöglichen den KI-Systemen bei neuen Fotos das passende Label vorherzusagen.

Beim unsupervised learning (Unüberwachtes Lernen) identifiziert das KI-System Muster in Datensätzen (Eingaben) und sortiert die Daten, die sich ähneln, in Gruppen (Ausgabe). So können zum Beispiel viele Bilder in Gruppen ähnlicher Bilder eingeteilt werden – entweder Bilder mit Katzen oder Bilder mit Hunden.

Beim reinforcement learning (Verstärkungslernen) lernt das KI-System durch das Feedback der Programmierenden. Dabei wird dem KI-System zurückgemeldet, ob seine Ausgabe richtig oder falsch war: „Ja, das Bild zeigt eine Katze" oder „Nein, das Bild zeigt keine Katze". Daraufhin verändert das KI-System sein Regelsystem, um möglichst viele richtige Ausgaben generieren zu können.[117]

Maschinelles Lernen ist hilfreich, wenn die Programmierenden nicht alle möglichen zukünftigen Situationen vorhersehen können oder wenn sie nicht wissen, wie sie selbst eine Lösung programmieren sollen. In solchen Fällen gibt es für die Maschine verschiedene Lernprobleme: Klassifizierung und Regression. Wenn der Output ein endlicher Satz von Werten sein soll, nennt sich das Klassifizierung, und wenn der Output eine Zahl sein soll, nennt sich das Regression.[118] Das geschieht mit Hilfe von verschiedenen mathematischen Modellen, wie zum Beispiel Entscheidungsbäumen oder linearen Regressionen. Je komplexer die Aufgabe für die KI-Systeme sind, desto komplexer werden auch die mathematischen Modelle,

116 Hume 2000[1739]; Russel, Norvig 2021:24, 670; Boddington 2023:3
117 Russel, Norvig 2021:671
118 Ebd.:669f.

welche das Erlernen eines passenden Regelsatzes ermöglichen sollen. Hier gibt es zum Beispiel Random Forests (Zufallswälder) oder Neuronale Netze, die so komplex sind, dass die Entstehung ihrer Regelsätze und auch die Regelsätze selbst für die Programmierenden nicht mehr nachvollziehbar sind.[119]

Für das Programmieren von Neuronalen Netzen wird eine hochkomplexe Technik verwendet: das Deep Learning, ein Teilgebiet des Machine Learning.[120] Das Deep Learning unterscheidet sich vom Machine Learning dadurch, dass die Berechnungswege von Eingabe zu Ausgabe viele, nicht nachvollziehbare Schritte umfassen. Hierdurch können auch unstrukturierte Daten wie Texte, Bilder, Töne und Videos verarbeitet werden.[121] Solche hochkomplexen KI-Systeme, die mit der Methode des Machine Learnings, oder auch konkreter mit Techniken des Deep Learnings, ihre eigenen Regelsysteme erstellen, sind nicht mehr interpretierbar. Russel und Norvig definieren Interpretierbarkeit folgendermaßen: „We say that a machine learning model is interpretable if you can inspect the actual model and understand why it got a particular answer for a given input, and how the answer would change when the input changes."[122] Wenn nicht mehr verständlich ist, warum das KI-System für eine bestimmte Eingabe eine bestimmte Ausgabe generiert und wie sich die Ausgabe ändern würde, wenn sich die Eingabe ändert, dann ist der erlernte Regelsatz unbekannt und das KI-System nicht interpretierbar. Will man trotzdem verstehen, warum das KI-System für eine bestimmte Eingabe eine bestimmte Ausgabe generiert und somit den erlernten Regelsatz durchschauen, muss man sich auf Methoden der Erklärbarkeit (Explainable AI; XAI) verlassen. Während eine Inspektion des eigentlichen KI-Systems Interpretierbarkeit ermöglicht, kann Erklärbarkeit nur durch einen separaten Prozess bereitgestellt werden, indem ein anderer Algorithmus zusammenfasst, wie das zu erklärende KI-System vorgeht.[123]

119 Russel, Norvig 2021:675–722
120 Ertel 2021:321; Russel, Norvig 2021:801ff.
121 Russel, Norvig 2021:801
122 Ebd.:729; Hervorhebung entfernt
123 Ebd.:729

Bei dieser autonomisierten Form der Datenverarbeitung mit Hilfe von KI-Systemen entstehen mehrere Risiken, die im Lauf der Arbeit vertieft erörtert werden. An dieser Stelle ist es wichtig festzuhalten, dass Wenn-Dann-Algorithmen bei gleichem Input immer den gleichen Output generieren – KI-Systeme hingegen können bei gleichem Input auch unterschiedlichen Output generieren, da sie mit jedem neuen Input ihr Regelwerk überarbeiten. Das deutet daraufhin, dass KI-Systeme nicht fehlerfrei sind – in Kombination mit der fehlenden Interpretierbarkeit von hochkomplexen Modellen wird das laut dem Informatiker Wolfgang Ertel zum Problem: „Die neue Qualität der Deep Learning Systeme(...) ist nun, dass es sehr schwer bis unmöglich wird, eine Stelle im System als Ursache des Fehlers zu identifizieren."[124] Das erschwert den Umgang mit fehlerhaften Ergebnissen erheblich und mahnt zur Vorsicht beim allgemeinen Umgang mit algorithmischen Ergebnissen. Denn eigentlich werden Fakten in Diskursen produziert, sodass ihre Entstehung am Ende nachvollziehbar ist.[125] Ein Algorithmus aber macht den Diskurs mit sich selbst aus und produziert nur eine Behauptung, deren Entstehung unter Umständen nicht nachvollziehbar ist.

Deshalb ist es wichtig, die Ergebnisse von Algorithmen nur als Sinnangebot zu verstehen, das noch eines Diskurses bedarf. Sie dürfen nicht als unstrittig verstanden werden, da es sich um gedeutete Daten handelt. Algorithmische Ergebnisse sind keine Tatsachen, da sie noch nicht gesellschaftlich akzeptiert werden konnten. Sie müssen sich als Sinnangebot noch einem gesellschaftlichen Diskurs stellen, wie jede menschliche Deutung von Daten auch. Wenn die Ausgaben eines Algorithmus als Fakten betrachtet werden, wird er zum Diskursakteur. Er darf dabei aber nicht mehr Macht haben als die Menschen, die nach der Entstehung dieser Fakten fragen. Der algorithmische Beitrag zum Diskurs kann gesellschaftliche Veränderungen anregen. Solch große Einflussmöglichkeiten erfordern die Übernahme von Verantwortung. Wer diese Verantwortung zu tragen hat, soll im Verlauf dieser Arbeit geklärt werden.

Eine Problemlösung für die Interpretierbarkeit der KI-Systeme und also die Identifikation von Fehlern können die Methoden der

124 Ertel 2021:343
125 Felder 2013:14ff.

Erklärbarkeit sein. So konnten zum Beispiel die Informatiker Marco Tulio Ribeiro, Sammer Singh und Carlos Guestrin zeigen, dass ein KI-System, welches Bilder von Huskys und von Wölfen unterscheiden kann, einen völlig falschen Regelsatz erlernt haben kann: „(...) the classifier predicts “Wolf" if there is snow (or light background at the bottom), and “Husky" otherwise, regardless of animal color, position, pose, etc.“[126] In diesem Fall wurde das KI-System absichtlich dazu gebracht, falsche Regeln zu erlernen, da die Trainingsdaten aus Bildern bestanden, welche nur dann Schnee im Hintergrund hatten, wenn Wölfe abgebildet waren, während Bilder von Huskys keinen Schnee im Hintergrund hatten. Aber solche Fehler können auch unabsichtlich geschehen und in Kontexten, in denen Menschen von den Ausgaben des KI-Systems betroffen sind, zu negativen Konsequenzen führen. So zum Beispiel, wenn eine Gesichtserkennung dunkelhäutige Frauen häufig falsch klassifiziert, hellhäutige Männer hingegen kaum. Die Informatikerinnen Joy Buolamwini und Timnit Gebru haben festgestellt, dass dies daran lag, dass die Trainingsdatensätze überwiegend aus Bildern von hellhäutigen Personen bestanden.[127] Auf dieser Grundlage hat das KI-System falsche Regeln erlernt.

Nicht immer ist so deutlich, warum das KI-System falsche Regeln erlernt hat oder nach welchen Regeln es überhaupt entscheidet. Eine einfache Erklärung dieser Regeln kann den bereits erwähnten Informatikern Stuart J. Russel und Peter Norvig zufolge zu einem falschen Gefühl der Sicherheit führen: „After all, we typically choose to use a machine learning model (rather than a hand-written traditional program) because the problem we are trying to solve is inherently complex, and we don’t know how to write a traditional program.”[128] Komplexe Probleme erfordern komplexe Lösungen und diese komplexen Lösungen lassen sich nicht unbedingt einfach erklären. Die selbsterlernten Regelsätze von KI-Systemen, die nicht interpretierbar sind, bleiben oft unverständlich. Trotzdem werden diese technischen Möglichkeiten verstärkt genutzt.[129]

126 Ribeiro et al. 2016:1142
127 Buolamwini, Gebru 2018
128 Russel, Norvig 2021:730
129 Ebd.

An dieser Stelle fragt sich, warum so viele Daten mit Hilfe von Wenn-Dann-Algorithmen und KI-Systemen verarbeitet werden. Shah et al. finden folgende Antwort darauf: „Data by itself has no value. However, the extraordinary value-adding potential of data lies in the ability to extract meaningful and actionable information from it."[130] Die algorithmische Analyse der Daten verspricht ein wertvolles Ergebnis. Wenn viele Daten kombiniert und unter ihnen Muster gefunden werden, kann dadurch neues Wissen entstehen. Wenn also zum Beispiel viele Daten über langjährige Pflanzenpflege gesammelt werden, machen diese in der Summe eine Aussage darüber, bei welcher Pflege welche Pflanze am längsten lebt. So entsteht Wissen darüber, welche Pflanzen viel Wasser brauchen und welche kein direktes Sonnenlicht vertragen – Daten über Pflanzenpflege versprechen Wissen darüber, wie Pflanzen am besten gepflegt werden können. So können Kausalitäten entdeckt werden, die zukünftig bessere Entscheidungen ermöglichen.

Da die Analyse von Daten sehr schnell zu Entscheidungsvorschlägen führen kann, sind digitale Daten zum wichtigsten Rohstoff geworden – sie werden sogar als das Öl des 21. Jahrhunderts bezeichnet.[131] Das Wertvolle an den Daten ist ihr Potenzial, durch ihre Interpretation neues Wissen zu erschaffen. Dieses neue Wissen ermöglicht dann Innovationen und damit auch finanzielle Vorteile. Wenn man zum Beispiel besser über die Pflege von Pflanzen Bescheid weiß, kann man den Ertrag pro Bodenfläche vergrößern und somit effektiver Landwirtschaft betreiben.

Dieses Wissen darf aber nicht unhinterfragt akzeptiert werden, weil Daten nicht neutral sind. Der Germanist Ekkehard Felder hingegen meint Daten seien „(...) unstrittige, allseits akzeptierte Fakten (...)."[132] So herrscht in der Literatur die Meinung vor, dass Daten in dem Sinne *gegeben* sind, dass sie überprüfbar sind, da sie durch Verhalten und Beobachtungen von Menschen entstanden sind. Distanzen sind messbar, Bewertungen von Schulleistungen sind nachweisbar etc. Jedes Datum für sich mag neutral sein, solange es überprüfbar ist und als deskriptive Aussage für sich steht. Auch Daten, die Aussagen über Menschen machen, sind erstmal deskriptiv. Aber

130 Shah et al. 2020:110
131 Streibich 2015:16
132 Felder 2013:14

wenn diese deskriptiven Aussagen für bare Münze genommen werden, kann es dadurch zu Diskriminierung kommen.

Diskriminieren bedeutet ursprünglich einfach nur unterscheiden (lat. discriminare), aber der Begriff wird seit dem 20. Jahrhundert mit einer negativen Wertung verbunden.[133] Dem Sozialpsychologen Andreas Zick zufolge lässt sich dies folgendermaßen erklären: „Diskriminierung beginnt bei der Wahrnehmung und endet in einer Herstellung und Etablierung von Ungleichwertigkeit, die Ungleichheit begründen soll."[134] Das heißt gemäß der Philosophin und Kommunikationswissenschaftlerin Larissa Krainer, dass Diskriminierung zwei verschiedene Prozesse beschreibt: Zum einen die Unterscheidung verschiedener Gruppen und zum anderen die unterschiedliche Bewertung dieser Gruppen. Letzteres funktioniert laut Krainer so: „Zur Abwertung von Gruppen werden häufig Stereotype entwickelt und angewandt."[135] Ein solches Stereotyp ist zum Beispiel, dass Frauen nicht mit Technik umgehen können. Dieser Stereotyp könnte dann ein Grund sein, Frauen als Gruppe abzuwerten im Kontext von Stellenbesetzungen bei einem Technologiekonzern wie zum Beispiel Amazon. Durch solche Stereotypen werden Ungleichheiten begründet und etabliert. Wenn solche Ungleichheiten Grundlage für Entscheidungen sind, können die Effekte dieser Entscheidungen als Diskriminierung – in dem negativ konnotierten Sinne – bezeichnet werden.[136] Diskriminierung bedeutet somit laut Informationswissenschaftlerin Batya Friedman und Technologiephilosophin Helen Nissenbaum eine Schlechterbehandlung einer Person oder Gruppe aus unvernünftigen oder unangemessenen Gründen.[137] Denn wenn es um die Frage geht, ob eine Person mit Technik umgehen kann und dementsprechend geeignet ist für einen Job bei Amazon, ist das Geschlecht irrelevant – somit wäre das Geschlecht der Bewerber:innen ein unvernünftiger Grund für ihre Ablehnung. Allein Faktoren wie das Interesse an Technik, die Ausbildung in dem Bereich oder Arbeitserfahrung sollten relevant und deshalb ein Grund sein, zwischen den Bewerber:innen zu unterscheiden.

133 Krainer 2024:304
134 Zick 2020:8
135 Krainer 2024:304
136 Mittelstadt et al. 2016:8
137 Friedman, Nissenbaum 1996:332

Die unhinterfragte Akzeptanz von Wissen, das aus Datensätzen gewonnen wurde, provoziert deshalb die Möglichkeit der Diskriminierung, weil bei der Analyse von Daten immer Gruppen voneinander unterschieden werden – dies ist die Grundlage von algorithmischer Mustererkennung. Die analysierten Datensätze bilden dabei nicht nur verschiedene Gruppen ab, sondern auch ihre Bewertung. Wenn in der Vergangenheit die Gruppe „Frauen" abgewertet und deshalb nicht eingestellt wurde, dann zeigt sich das in den Daten über bisher erfolgreiche Bewerber:innen. Diese Daten machen eine deskriptive Aussage darüber, welche Kandidat:innen (bzw. meistens Kandidaten) in der Vergangenheit im Bewerbungsprozess erfolgreich waren. Werden diese Daten aber als Grundlage für zukünftige Entscheidungen über die besten Bewerber:innen verwendet, gehen auch alle bisherigen Abwertungen von Gruppen als Status quo ein – so werden Diskriminierungen reproduziert.

Auf diese Weise bilden Datensätze Strukturen ab, die in der Gesellschaft vorherrschen und diskriminierend sein können. Auch wenn ein vermeintlich objektiver Algorithmus die Daten analysiert, findet er in repräsentativen Daten die Muster, die auch in der Gesellschaft zu finden sind. Dabei kann es sich um sexistische Strukturen handeln, weil zum Beispiel bei Amazon weniger Frauen als Männer arbeiten. Bekommt der Algorithmus diese Daten gefüttert, werden seine Ergebnisse dieses Muster repräsentieren. So könnte ein Algorithmus, der dabei helfen soll, die besten Bewerber:innen zu finden, Frauen systematisch ausschließen.[138] Ebenso machen sich rassistische Strukturen bemerkbar, wenn ein Algorithmus bei der Bilderkennung helle Gesichter besser erkennen kann als dunkle Gesichter.[139]

Im Falle der fehlenden Akzeptanz von weiblichen Bewerberinnen bei Amazon reproduziert sich eine gesellschaftliche Ungleichheit und zugeschriebene Ungleichwertigkeit, weil die Daten die Wirklichkeit abbilden. Im Gegensatz dazu entsteht bei der Gesichtserkennung eine Ungerechtigkeit dadurch, dass die Trainingsdaten unvollständig waren und die Gesellschaft nicht vollständig abgebildet haben. Bei der Verarbeitung von Daten gibt es also zwei Möglichkeiten

138 Stahl et al. 2023:10f.
139 Buolamwini, Gebru 2018

der Verzerrungen (Bias).[140] Es besteht zum einen die Gefahr, dass die Daten nicht die Wirklichkeit abbilden und das Ergebnis somit keine Aussage über die Wirklichkeit machen kann. Zum anderen besteht die Gefahr, dass die Daten eine ungerechte Wirklichkeit abbilden, die sich durch ihre Analyse in den Ergebnissen reproduziert.

Wenn man in diesem Kontext die ethische Frage stellt, ob bestimmte Daten zu einem bestimmten Zweck verarbeitet werden sollten, muss zwischen Nutzen und Risiken abgewogen werden. Wenn beispielsweise die Gesundheitsdaten von Personen mit einer chronischen Krankheit analysiert werden, kann womöglich Wissen über die Entstehung der Krankheit oder über die bestmögliche Behandlung extrahiert werden. Denn vielleicht hat sich die Mehrheit der erkrankten Menschen zuvor ungesund ernährt, oder ein Medikament schlägt bei jungen Menschen besser an als bei älteren, aber in Kombination mit einer bestimmten Diät ist es auch bei älteren Menschen wirksam. So kann Prävention gegen die Erkrankung betrieben werden, und es kann die beste Behandlung für alle Erkrankten gefunden werden – die Vorteile der Datenverarbeitung sind hier deutlich zu erkennen. Für die bereits Erkrankten, deren Daten analysiert werden, besteht jedoch durch die Datenverarbeitung das Risiko, dass Unbefugte Zugriff auf ihre Daten bekommen. Denn die Information, dass man eine chronische Krankheit hat, kann zu Schlechterstellung führen, wenn zum Beispiel ein Arbeitgeber die geplante Festanstellung der erkrankten Person überdenkt. Die Risiken, die im Datenökosystem entstehen, werden in Kapitel 3.3 identifiziert und in den Kapiteln 6 und 7 systematisch betrachtet.

Bis hierhin ist erstmal klar, dass das Sammeln und Verarbeiten von Daten sowohl Vorteile hat, als auch Risiken bereithält. Da in dieser Arbeit erörtert werden soll, wie Datenökosysteme ethisch bewertet und dementsprechend gestaltet werden sollten, wird nun definiert, was ein Datenökosystem ist.

140 Eine ausführliche Definition von „Bias“ folgt in Kapitel 7 mit drei verschiedenen Formen der Verzerrung nach Friedman und Nissenbaum (1996).

2.2 Definitionen von Datenökosystemen

Ein Datenökosystem lässt sich als System definieren, das sich zum Ziel setzt Daten zu sammeln und auszuwerten. Der Logiker und Sprachphilosoph Stewart Shapiro definiert jegliches System als „(...) a collection of objects with certain relations."[141] Bei einem Datenökosystem handelt sich also um eine Sammlung von Objekten zwischen denen Daten fließen. In diesem Teilkapitel wird gezeigt, dass die Beziehungen zwischen den Objekten den Datenfluss bestimmen.

Bevor die Idee eines Datenökosystems entstand, war die Nutzung von Daten eine Einbahnstraße, da es keine Rückkopplungsschleife gab zwischen denen, die die Daten bereitgestellt haben und denen, die sie genutzt haben.[142] Dieses Phänomen der Rückkopplung, das ein Datenökosystem von anderweitiger Datenverarbeitung abhebt, wird ausgedrückt durch den Begriff „Ökosystem". Als Mitarbeitende des Bundesministeriums für Wirtschaft und Energie schrieben Thomas Bendig und seine Kolleg:innen der Begriff Ökosystem „(...) greift die Metapher natürlicher Ökosysteme als definierten Lebensraum vielfältiger Organismen und ihrer Umwelt auf."[143] Laut Antti Poikola, der sich als Vorreiter bei der Förderung offen zugänglicher öffentlicher Daten in Finnland bezeichnet, und den Forschern und Designern Petri Kola und Kari A. Hintikka bezieht sich der Begriff des Ökosystems auf ein funktionierendes Ganzes in einem bestimmten Bereich.[144] Konkret funktioniert dieses Ganze als eine zyklische, biologische Umgebung. Laut dem Betriebswirtschaftler Maximilian Heimstädt und seinen Kollegen Fredric Saunderson und Tom Heath, die im Datenschutz Bereich arbeiten, bilden im digitalen Ökosystem die Datensätze sowie die Systeme und Akteure, die diese Daten unterstützen, eine Analogie dazu.[145]

Die Rückkopplung kann sich im Datenökosystem zeigen, indem Datenquellen jedes Mal auch von der Verarbeitung ihrer Daten profitieren, sodass immer ein einmaliger Kreislauf des Datenflusses

141 Shapiro 1997:73
142 Oliveira et al. 2019:590; Pollock 2011
143 Bendig et al. 2019:5; Stand 2025 ist Thomas Bendig Mitarbeiter beim Bundesministerium für Digitales und Staatsmodernisierung.
144 Poikola et al. 2011:13
145 Heimstädt et al. 2014:4

entsteht. Wenn allerdings nicht nur der Zyklus des Prozesses der Datenverarbeitung gemeint ist, sondern auch die zyklische Verwendung von Daten als Ressourcen gefordert wird, stößt die Metapher des Ökosystems bei der Verwendung personenbezogener Daten an rechtliche Grenzen. Denn laut Datenschutzgrundverordnung müssen die Datenquellen, über welche die Daten eine Aussage machen, im Rahmen der Einverständniserklärung zur Verarbeitung ihrer Daten über die Zwecke der Verarbeitung informiert werden.[146] Das heißt, personenbezogene Daten dürfen nicht unangekündigt für andere Zwecke wiederverwendet werden, was den zirkulären Ressourcenfluss erschwert – er müsste vor dem Sammeln der Daten detailliert geplant werden.

Schon bei der Begriffserklärung wird der erste Konflikt deutlich, denn die Idee, dass bereits verwendete Daten in bereinigter, integrierter und verpackter Form zur Wiederverwendung in das Datenökosystem eingebracht werden sollen[147], steht bei der Verwendung von personenbezogenen Daten den Interessen der Personen entgegen. Der eigentliche Sinn eines Datenökosystems ist es aber, dass alle Beteiligten von der Arbeit mit den Daten profitieren.[148] So wird deutlich, dass zwischen den Akteuren Kompromisse erforderlich sind. Das passt aber tatsächlich in den natürlichen Verlauf, da Datenökosysteme sowieso erst durch eine fruchtbare Interaktion zwischen kooperierenden und konkurrierenden Akteuren entstehen.[149] Laut dem Wirtschaftswissenschaftler Ron Adner kann ein Ökosystem entweder als Gemeinschaft von Akteuren oder als Struktur von Aktivitäten, die sich durch ihr Wertversprechen definieren, verstanden werden.[150] In beiden Fällen ist eine koordinierte Zusammenarbeit notwendig, um gemeinsam einen Mehrwert aus den Daten zu ziehen. Das Ziel dieser Zusammenarbeit ist laut Poikola et al. die Verteilung und Entwicklung des Rohstoffes, also der Daten.[151] Ein Datenökosystem besteht demnach aus mehreren vernetzten Akteuren, die Daten wie einen Rohstoff betrachten, der sich idealerwei-

146 Artikel 5, Absatz 1 (b) DSGVO
147 Pollock 2011
148 Poikola et al. 2011:7
149 Heimstädt et al. 2014:1
150 Adner 2017:40
151 Poikola et al. 2011:13

se in einem Kreislauf befindet, dadurch das Ökosystem speist, zur Wertschöpfung beiträgt und allen Beteiligten Vorteile bringt.[152]

Laut Adner spielen insbesondere die Akteure im Datenökosystem eine große Rolle: „The ecosystem is defined by the alignment structure of the multilateral set of partners that need to interact in order for a focal value proposition to materialize."[153] Die Ausrichtungsstruktur bezeichnet das Ausmaß, in dem sich die Akteure über ihre Positionen im System einig sind. Erfolgreich sei ein Ökosystem dann, wenn alle Akteure mit ihren Positionen zufrieden seien. Dabei handele es sich um eine Vielzahl von Akteuren mit multilateralen Beziehungen, die sich nicht auf bilaterale Interaktionen herunterbrechen ließen.[154] Die am System beteiligten Akteure haben die gemeinsame Wertschöpfung zum Ziel. Die Akteure, von deren Beteiligung die geplante Wertschöpfung abhängt, bezeichnet Adner als Partner. Um die Wertschöpfung verwirklichen zu können, werden die dafür erforderlichen Aktivitäten durch das Einbeziehen und Koordinieren von Partnern durchgeführt. Laut Adner ist die gemeinsame Wertschöpfung dabei die Grundlage des Ökosystems.[155]

Adners Definition liegt eine aktivitätsbasierte Perspektive auf die Interdependenzen im Ökosystem zugrunde, die er die Ökosystem-als-Struktur Perspektive nennt[156]: „(…) the ecosystem-as-structure view begins [with] the value proposition, considers the activities required for its materialization, and ends with actors that need to be aligned (…)."[157] Aus diesem Blickwinkel beginnt ein Ökosystem also mit der Absicht einer Wertschöpfung und strebt das Identifizieren der Gruppe von Akteuren an, die interagieren müssen, um das Vorhaben zu verwirklichen.[158]

Eine andere mögliche Perspektive wäre laut Adner der akteurszentrierte Ökosystem-als-Zugehörigkeit Ansatz: „The ecosystem-as-affiliation approach begins with the actors (usually defined by their ties to a focal actor), considers the links among them, and ends with the possible value propositions and enhancements that the ecosys-

152 Zubcoff et al. 2016:251f.
153 Adner 2017:40, 42
154 Adner 2017:42
155 Ebd.:43
156 Ebd.:41
157 Ebd.:44
158 Ebd.:41

tem can generate."[159] Hierbei stehen die Akteure eines Ökosystems im Mittelpunkt, aber es werden nicht ihre Aktivitäten betrachtet. Dieser Ansatz liegt zum Beispiel der Definition eines staatlichen Ökosystems von der Kommunikationswissenschaftlerin Teresa M. Harrison und ihren Kolleginnen Theresa A. Pardo und Meghan Cook aus dem Center for Technology in Government zugrunde: „Our perspective on the *open government ecosystem* envisions government organizations as central actors, taking the initiative within networked systems organized to achieve specific goals related to innovation and good government."[160] Allerdings dürfen die Aktivitäten der Akteure meiner Meinung nach in der Definition nicht fehlen, weil Akteure sich durch ihre Aktivitäten auszeichnen und ohne diese auch keine Wertschöpfung betreiben können.

Unabhängig davon, ob Ökosysteme aus der Struktur- oder Zugehörigkeits-Perspektive betrachtet werden, spielen die Akteure eine wichtige Rolle, da die Wertschöpfung im Datenökosystem in einem System von Akteuren stattfindet.[161] Poikola et al. stimmen zu: „The ecosystem sees organisations and private people as actors, interacting with each other to mould the conventions of [the] ecosystem."[162] Die Elemente eines Datenökosystems sind Ressourcen, Aktivitäten, Akteure, Rollen, Positionen im Fluss der Aktivitäten und Beziehungen zwischen den Akteuren.[163] Die Informatiker:innen Marcelo Iury S. Oliveira, Glória de Fátima Barros Lima und Bernadette Farias Lóscio beschreiben die Akteure eines Datenökosystems folgendermaßen: „Actors are autonomous entities such as enterprises, institutions, or individuals, which act in the Data Ecosystem in one or more specific roles."[164] Die Akteure werden durch eine Reihe von Interessen motiviert und haben in der Regel eine Verpflichtung gegenüber dem Datenökosystem – dadurch können sie einen Anreiz haben für ihre Aktivitäten.[165] Jeder Akteur hat eine Rolle in den Informations-, Material- und Geldflüssen und diese beeinflussen die

159 Adner 2017:43f.
160 Harrison et al. 2012:907
161 Immonen et al. 2014:88
162 Poikola et al. 2011:28
163 Adner 2017:43f.; Oliveira, Lóscio 2018:8
164 Oliveira et al. 2019:606
165 Ebd.

Beziehungen untereinander.[166] Oliveira et al. definieren den Begriff der Rolle so: „A role is a function played by an actor in the ecosystem. It is related to a position and a set of duties."[167] Gewisse Pflichten entstehen aus der spezifischen Rolle und Position des Akteurs im Datenökosystem.

Ein Datenökosystem ist also ein soziotechnisches System, das gemeinsam die Verarbeitung von Daten anstrebt, um einen Mehrwert aus den Daten zu ziehen.[168] Dieses soziotechnische System besteht aus Menschen und Organisationen einerseits, Infrastruktur und Technologie andererseits, und seine Inhalte und Vorgehensweisen werden durch Ressourcen und Richtlinien bestimmt. Die wichtigsten Ressourcen in einem Datenökosystem sind dabei Datensätze.[169] Zu der Verarbeitung von Daten gehören Sammlung, Integration, Analyse, Speicherung, gemeinsame Nutzung, Datensicherheit und -schutz sowie Archivierung. Ein Mehrwert wird aus den Daten gezogen, wenn durch ihre Verarbeitung zum Beispiel Dienstleistungen besser erbracht werden können und wenn dadurch eine Politik geschaffen wird, die Bürgern, Unternehmen und staatlichen Stellen zugutekommt.[170] Diese erforschende, wertschöpfende Verarbeitung von Daten ist nur möglich durch die Zusammenarbeit der verschiedenen Akteure im Datenökosystem. Diese Akteure – der soziale Teil des soziotechnischen Systems – sind für das Produzieren, Bereitstellen oder Konsumieren der Daten zuständig, und sie sind dabei durch Beziehungen miteinander verbunden.[171] In der Literatur finden sich unterschiedliche Definitionen von Datenökosystemen, aber die Konzepte der Akteure, Rollen, Beziehungen und Ressourcen sind allen gemeinsam.[172]

Ein weiterer interessanter Begriff ist der des Open Data Ecosystems. Der Begriff „Open Data" bezieht sich auf die offene Zugänglichkeit der Daten innerhalb eines Datenökosystems und wird von der Open Knowledge Foundation, eine Stiftung, die sich für digital allen Menschen zugängliches Wissen einsetzt, folgendermaßen defi-

166 Immonen et al. 2014:88
167 Oliveira et al. 2019:606
168 Shah et al. 2020:112; Oliveira et al. 2019:590
169 Oliveira, Luscio 2018:7
170 Shah et al. 2020:112
171 Oliveira et al. 2019:604
172 Ebd.:604

niert: „Open data and content can be freely used, modified, and shared by anyone for any purpose".[173] Die Offenheit der Daten ist dabei nicht als ihre inhaltliche Transparenz zu verstehen, sondern allein als unbegrenzter Zugang zu den Daten.[174] Bei Open Data Ecosystems besteht eine der Leistungen darin, die Daten öffentlich bereitzustellen. Hier passt die Beschreibung von sogenannten „Data spaces" (Datenräume) aus dem Data Governance Act, eine Verordnung der Europäischen Union zur Erleichterung der gemeinsamen Nutzung und Wiederverwendung öffentlich zugänglicher, geschützter Daten über Sektoren und EU-Länder hinweg. Diese Datenräume sind definiert als Datenökosysteme in einem bestimmten Anwendungsbereich, die auf gemeinsamen Regeln basieren. Solche Datenräume enthalten Daten und ermöglichen allen Nutzenden den Zugriff darauf. Dabei werden von einem Akteur mehrere Datenökosysteme gleichzeitig genutzt, sodass sie sich in und zwischen verschiedenen Datenräumen bewegen.[175]

In der Idealvorstellung von Datenräumen haben die Akteure die Kontrolle über ihre eigenen Daten. So schreiben der CEO von der International Data Spaces Association, Lars Nagel, und der Senior Vizepräsident von Innopay, Douwe Lycklama: „Data will only be shared and exchanged between actors if they decide to do so."[176] Daten können zum Beispiel geteilt werden, weil sich die Akteure davon einen finanziellen Gewinn versprechen. Aber es muss dabei nicht um eine finanzielle Transaktion gehen.[177] Es kann auch mit Daten gehandelt werden, wenn diese für einen Akteur im Ökosystem einen Wert haben oder um die Qualität eines Produktes bzw. einer Dienstleistung zu verbessern.

An einigen Stellen wird Open Data gepriesen, weil sich der offene Zugang zu Daten positiv auf die Wertschöpfung im Datenökosystem auswirkt, weil er Bürgerbeteiligung sowie Innovation fördert, weil er das Wirtschaftswachstum ankurbelt, weil er die Bereitstellung öffentlicher Dienstleistungen verbessert und dabei Kosten spart.[178] Die

173 Open Knowledge Foundation 2022; Hervorhebungen entfernt
174 Yu, Robinson 2012:208
175 Nagel, Lycklama 2021:7
176 Ebd.:92
177 Steffen et al. 2021:46
178 Zuiderwijk et al. 2016:223; Zuiderwijk et al. 2014:17; Davies 2011:1; Zeleti, Ojo 2014:477

Offenheit von Daten im Allgemeinen wird sogar gefordert, zum Beispiel von dem Informatiker Mark A. Parsons und seinen Kollegen: „Data should generally be openly accessible."[179] Parsons et al. versprechen sich davon, dass die Interaktion zwischen den Akteuren im Datenökosystem dadurch gefördert wird und dass infolgedessen die Zusammenarbeit und das interdisziplinäre Verständnis verbessert werden.[180] Manche Autor:innen argumentieren, dass der öffentliche Zugang zu Daten notwendig ist, um die Transparenz der Regierung zu ermöglichen[181] und dass Open Data dadurch auch eine stärkere demokratische Rechenschaftspflicht möglich macht[182]. Hierbei würde es sich dann um ein sogenanntes Open Government handeln, das durch frei zugängliche, wiederverwendbare und weiterverbreitete Daten seinen eigenen Zielen gerecht werden kann.[183]

Allerdings kommt der Sozialwissenschaftler Tim Davies zu dem Schluss, dass Open Data vor allem gut organisierte Interaktion zwischen den Akteuren im Datenökosystem fordert. Gute Zusammenarbeit im Rahmen einer stabilen Infrastruktur entsteht nicht von alleine, sondern müsse mit Hilfe von Koordinierungsstrategien erarbeitet werden.[184] Außerdem macht der Begriff Open Data an sich keine Aussage über den Inhalt der Daten, sondern nur über ihren technischen und rechtlichen Zustand.[185] Deshalb kann allein aus dieser technischen Beschreibung der Zugangsmöglichkeiten zu Daten noch kein Schluss gezogen werden über den tatsächlichen oder erwarteten Nutzen der Datenweitergabe.[186]

Datenökosysteme entstehen auch in der deutschen öffentlichen Verwaltung. Denn das Onlinezugangsgesetz (OZG) verpflichtet die deutsche Verwaltung dazu, ihre Services auch in digitaler Form anzubieten. Laut einem Bericht des Bundesministeriums für Wirtschaft und Energie (BMWi) über das Projekt Gaia-X setzt ein solcher digitaler Wandel „(...) eine leistungsfähige Cloud-Infrastruktur voraus

179 Parsons et al. 2011:557
180 Ebd.:565
181 Zuiderwijk et al. 2014:17; Geiger, Lucke 2012:265; Zubcoff et al. 2016:250
182 Davies 2011:1; Zeleti, Ojo 2014:477
183 Zubcoff et al. 2016:250
184 Davies 2011:5
185 Heimstädt et al. 2014:2; Yu, Robinson 2012:188f., 206
186 Yu, Robinson 2012:208

(...)."[187] Durch das europäische Projekt Gaia-X soll eine vernetzte Dateninfrastruktur geschaffen werden. Zu diesem Zweck werden verschiedene Projekte durch das Bundesministerium für Wirtschaft und Klimaschutz gefördert – darunter auch das in der Einleitung erwähnte MERLOT Projekt.[188]

Ziel dieser Datenökosysteme ist es, das Teilen von Informationen so leicht zu machen, dass die Dienstleistungen der Verwaltung schneller erbracht werden können und mit weniger Aufwand für die Antragstellenden verbunden sind. Bei der Beantragung von Elterngeld muss beispielsweise der Einkommensnachweis der Eltern eingereicht werden.[189] Damit die Eltern nicht ihre Einkommensnachweise suchen und damit zu einem Rathaus gehen müssen, sollen die Einkommensnachweise aus anderen Kontexten genutzt werden können. Wenn die Eltern vorher schon Bürgergeld oder Bafög beantragt haben, mussten sie dafür ebenfalls ihr Einkommen angeben. Das Datenökosystem soll ermöglichen, dass die antragstellenden Eltern der Elterngeldbehörde den Zugriff auf ihre Einkommensnachweise aus dem Bürgergeld- oder Bafög-Antrag ermöglichen. Dadurch würden die Daten der Antragstellenden (Datenquellen) zu verschiedenen Zwecken verarbeitet, und sowohl die Antragstellenden selbst als auch die bearbeitenden Behörden profitieren davon durch einen geringeren Aufwand und Zeitersparnis.

Der Begriff „open data" könnte sich im Kontext der öffentlichen Verwaltung zum einen darauf beziehen, dass die Daten des Ökosystems allgemein zugänglich sind oder zum anderen auf die Bereitstellung der Daten für den öffentlichen Sektor, sodass die Daten verwaltungsintern zugänglich sind. Die Daten, die allgemein zugänglich gemacht werden könnten, wären Daten aus der Verwaltung oder aus den verschiedenen Ministerien und Ämtern. Auf diese Weise könnten Daten aus verschiedenen Bereichen miteinander verknüpft werden. So könnten zum Beispiel Daten über das Einkommen einer Person aus dem Bürgergeldantrag beim Bundesarbeitsministerium weitergeleitet werden an das Bundesministerium für Wohnen, Stadtentwicklung und Bauwesen für den Wohngeldantrag der gleichen Person.

187 Bundesministerium für Wirtschaft und Energie 2019:34

188 Mehr Informationen unter: https://www.bmwk.de/Redaktion/DE/Dossier/gaia-x.html

189 Bundesministerium der Finanzen 2023

Daraufhin stellt sich die Frage, ob auf Grund des erwarteten Nutzens der Datenweitergabe in Datenökosystemen alle Daten als Open Data öffentlich zugänglich sein sollten. Die erwarteten positiven Auswirkungen von offen zugänglichen Daten reichen von einer Steigerung der Wertschöpfung im Datenökosystem über eine Verbesserung von Dienstleistungen und Förderung von Bildung bis hin zur Schaffung neuer Innovationen und Erreichung von mehr Transparenz in der Demokratie.[190] Sobald es sich aber um personenbezogene Daten handelt, wird es eine Abwägung zwischen den Vorteilen und dem Schutz der Privatsphäre der Datenquellen geben müssen.

Abschließend kann man sagen, dass sich ein Datenökosystem über den Datenfluss zwischen Akteuren mit verschiedenen Rollen definiert. Zudem wurde wieder deutlich, dass insbesondere der offene Umgang mit digitalen Daten viele Vorteile verspricht. Es wurde aber auch angedeutet, dass dieser Umgang mit den Daten für bestimmte Akteure auch Risiken bereithält.

2.3 Akteure im Datenökosystem

Um ein besseres Verständnis von den Akteuren im Datenökosystem zu bekommen, stelle ich ein eigenes Modell von der Struktur eines Datenökosystems auf. Die Struktur definiert sich dabei laut Shapiro wie folgt: „A *structure* is the abstract form of a system, highlighting the interrelationships among the objects, and ignoring any features of them that do not affect how they relate to other objects in the system."[191] Das heißt, das Modell stellt nur die abstrakte Form eines Datenökosystems dar, indem die Beziehungen zwischen den Objekten hervorgehoben werden. Die Objekte sind hier die menschlichen Akteure, da die technologischen Objekte des Datenökosystems selbst keine Beziehungen eingehen können – sie werden nur benutzt durch die menschlichen Akteure. In diesem Modell der Struktur eines Datenökosystems werden also nur die Merkmale der Objekte dargestellt, die einen Einfluss darauf haben, wie sie sich zu anderen Objekten im System verhalten.

190 Yu, Robinson 2012:182; Poikola et al. 2011:11, 14; Shah et al. 2020:112
191 Shapiro 1997:74

Das Modell von der Struktur eines Datenökosystems soll auf Grundlage der unter 2.2 erfolgten Definition folgende Kriterien erfüllen:

1) Der Datenfluss soll klar erkennbar sein.
2) Die beteiligten Akteure sollen voneinander differenziert sein.
3) Die Rollen der Akteure sollen klar ersichtlich sein.

Es gibt bereits Modelle von Datenökosystemen, doch sie alle erscheinen mir unübersichtlich oder sogar unvollständig. Entweder fehlt die Bezeichnung der Akteure, sodass unklar ist, wer die Aufgaben im Datenökosystem ausführt[192], oder essentielle Schritte – wie die Datenverarbeitung – werden nicht separat betrachtet, sondern zum Beispiel den Nutzenden als Aufgabe zugeschrieben[193]. In einem anderen Modell ist es jedem Menschen möglich einen eigenen Datensatz oder Algorithmus zur Verfügung zu stellen[194], was das Überprüfen der Qualität erschweren würde.

Suwook Ha, Seungyun Lee und Kangchan Lee, die am Electronics and Telecommunications Research Institute forschen, haben 2014 ein sehr übersichtliches Modell erstellt mit Datenprovidern, Serviceprovidern und Datenkonsumierenden, welches in Abbildung 1 zu sehen ist.

Die Aufgabe der Datenprovider ist es, neue Datensammlungen in das Datenökosystem einzuführen, damit diese verarbeitet werden können. Der Serviceprovider bietet die dafür notwendigen Hilfsmittel an: Infrastrukturen für den Verbraucher und Analyseanwendungen. Die Analyseanwendungen ermöglichen Datenvisualisierung und -analyse, und die Infrastruktur ermöglicht die Sammlung, Speicherung und Vorverarbeitung der Daten. Die Datenkonsumierenden nutzen die Ergebnisse der Datenanalyse und können diese außerhalb vom Datenökosystem zur Verfügung stellen.[195] Dieses Modell zeigt den Verlauf der Daten innerhalb des Verarbeitungsprozesses klar auf (Kriterium 1), es wird jedoch nicht erkennbar, woher die Daten kommen. Das Fehlen der Datenquellen als Akteur halte ich für ein Problem, weil dadurch der Schutz der Datenquellen bei den Abläufen in einem Datenökosystem keine Rolle zu spielen scheint.

192 Davies 2011:3; Attard et al. 2016:454f.
193 Dawes et al. 2016:39
194 Traub et al. 2020:8
195 Ha et al. 2014:166

Es ist so nicht möglich, die Position der Datenquellen zu problematisieren.

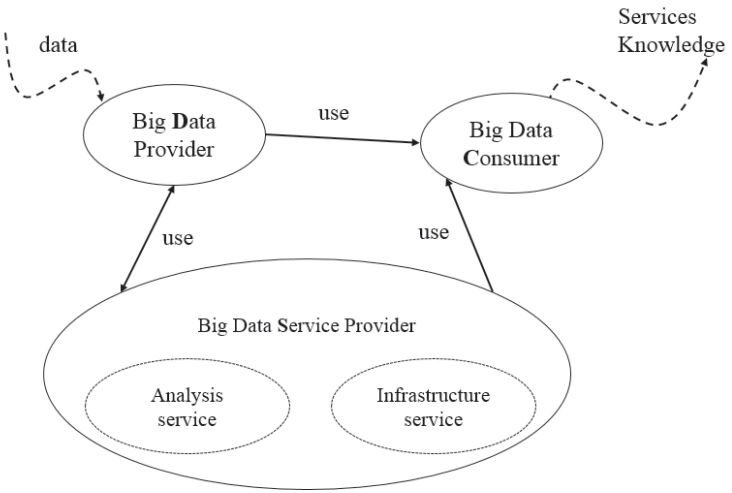

Abbildung 1 Big Data Ökosystem von Ha et al.[196]

Das gleiche Defizit haben die bereits erwähnten Technologieforscher Syed Iftikhar Hussain Shah, Vasilis Peristeras und Ioannis Magnisalis in ihrem Modell aus dem Jahr 2020. Sie beschreiben vier „Elemente" eines Datenökosystems und identifizieren ihre jeweilige Rolle und Motivation.[197] Ich würde sagen Shah et al. beschreiben nicht die Elemente eines Datenökosystems (Ressourcen, Aktivitäten, Akteure, Rollen, Positionen im Fluss der Aktivitäten und Beziehungen zwischen den Akteuren[198]), sondern sie beschreiben vier Akteure eines Datenökosystems. Demnach sind die vier Akteure eines Datenökosystems *Data Publishers*, die Daten kostenlos bereitstellen oder verkaufen, eine *Data Business Entity*, welche sowohl eigene Daten beisteuert und diese analysiert als auch die beste Datenquelle für die Bedürfnisse der Data User findet, *Data Support Serviceprovider*, die

196 Ha et al. 2014:166
197 Shah et al. 2020:106ff.
198 Adner 2017:43f.; Oliveira, Lóscio 2018:8

gegen Geld einen Speicherort für die Daten bereitstellen, und *Data Users*, die dank der Analyse der Daten Wert daraus schöpfen.[199] Eine der Aufgaben der Data Business Entity ist es zwar, bestmögliche Datenquellen zu finden, dabei sind aber die Datenquellen kein Akteur des Datenökosystems. Aus meiner Sicht erfüllen die Modelle von Ha et al. und Shah et al. dadurch weder Kriterium 2 noch 3.

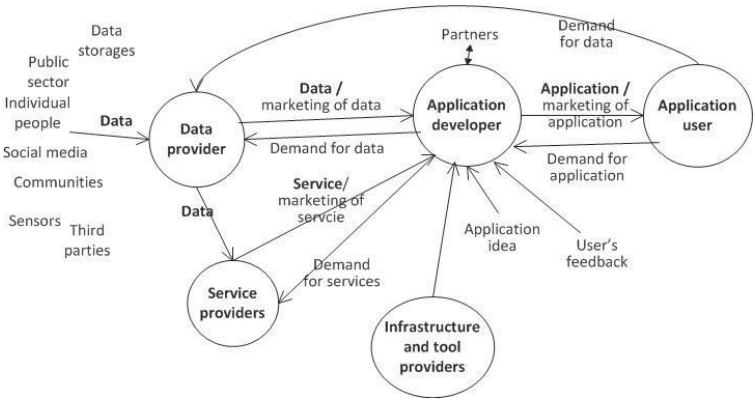

Abbildung 2 Offenes Datenökosystem von Immonen et al.[200]

Anders ist das bei Anne Immonen, Marko Palviainen und Eila Ovaska aus dem VTT Technical Research Centre of Finland – sie schließen in ihrem Modell Datenquellen ein, wie in Abbildung 2 erkennbar ist. Die Daten gelangen zum Beispiel von Individuen oder Sozialen Medien zu den Datenprovidern. Diese stellen die Daten wiederum Serviceprovidern zur Verfügung, sowie denjenigen, die die Anwendungen entwickeln (Application developer). Die Serviceprovider produzieren daraufhin datenbezogene Dienste. Mit Hilfe dieser Dienste und angebotener Infrastruktur entwickeln die Application developer Anwendungen für die Daten. Diese werden von den Anwendungsnutzenden (Application user) konsumiert. Das heißt, die Anwendungen nutzen Daten und Dienste und produzieren wertvolle Informationen und Wissen für die Nutzenden.[201] Obwohl dieses Modell in seiner Vollständigkeit überzeugt, erscheint

199 Shah et al. 2020:106–109
200 Immonen et al. 2014:91
201 Ebd.:91

mir der Datenfluss über die Dienstanbietenden unnötig kompliziert. Da Dienst und Anwendung nicht separat entwickelt werden müssen, ist es problemlos möglich, den Datenfluss im Modell zu vereinfachen. Das Modell von Immonen et al. erfüllt also Kriterium 2 und 3, da die Akteure und ihre Rollen klar erkennbar sind. Kriterium 1 wird jedoch nicht erfüllt, weil der Datenfluss sehr unübersichtlich abgebildet ist.

In der Literatur kommen viele interessante Ideen zusammen, sie vermischen sich allerdings auch – hier muss begriffliche und definitorische Klarheit geschaffen werden. Typische Rollen, die in den meisten Modellen eines Datenökosystems auftauchen, sind Datenproduzierende, Datenanbietende und Datennutzende.[202] An dieser Stelle sind die Begrifflichkeiten schon missverständlich. Sind Datenproduzierende die Quellen der Daten oder diejenigen, die Daten zusammentragen und in diesem Sinne Datensätze produzieren? Bieten Datenanbietende die Daten zur Analyse an oder zur Nutzung nach der Verarbeitung? Aus meiner Sicht wird es klarer, wenn man von den Akteuren *Datenquellen*, *Datengenerierenden*, *Datenverarbeitenden* und *Datennutzenden* spricht. Dabei handelt es sich um Personen, die ihre Daten zur Verfügung stellen (*Datenquellen*), verschiedene Beteiligte, die die Daten sammeln (*Datengenerierende*), diejenigen, die die Daten kombinieren und analysieren, um sie dann aufbereitet zur Verfügung stellen zu können (*Datenverarbeitende*) und die Personen, die mit den aufbereiteten Daten arbeiten (*Datennutzende*). Diese Rollen beschreiben zugleich die vier wichtigsten Akteure eines Datenökosystems, weil sie alle direkt am Datenfluss beteiligt sind.

Welche Aktivitäten von welchen Akteuren ausgeführt werden, ist in der Literatur nicht eindeutig und wird deshalb an dieser Stelle erörtert. Laut einigen Modellen gibt es einen Keystone Actor – einen Akteur, der die Schlüsselrolle spielt, weil er für Stabilität sorgt und eine treibende Kraft im Datenökosystem darstellt.[203] In vielen Modellen wird er durch die Regierung repräsentiert, dabei handelt es sich dann um sogenannte Open Government Data Ecosystems.[204] Keystone Actors sind laut den Informatikerinnen Bonnie A. Nardi

202 Oliveira et al. 2019:606
203 Ebd.:608
204 Oliveira, Loscio 2018:7

und Vicki L. O'Day „(...) skilled people whose presence is necessary to support the effective use of technology."[205] Diese Menschen fungieren als Übersetzende, Moderierende, Lehrende und Vermittelnde, indem sie verschiedene Institutionen und unterschiedliche Disziplinen zusammenbringen.[206] Der keystone actor verwaltet also das Datenökosystem und sollte laut Oliveira et al. deshalb formale Richtlinien, Regeln und Praktiken fördern sowie Verantwortung für Überwachung, Bewertung, Entscheidungsfindung und das Ergreifen von Maßnahmen übernehmen.[207] Nach der Ansicht der Technologieforscher:innen François van Schalkwyk, Michelle Willmers und Maurice McNaughton ermöglicht der keystone actor den Datenfluss durch seine Vermittlung zwischen Institutionen und Disziplinen – aber er ist keine notwendige, treibende Kraft: „Keystone species are enablers, not necessarily drivers in the ecosystem; they can be useful but they are not essential to the sustained functioning of an ecosystem."[208] Ich nenne diejenigen, die im Datenökosystem zwischen Institutionen und Disziplinen vermitteln sowie Verantwortung tragen für Regeln, Überwachung und Maßnahmen, im Folgenden die Verwaltenden. Dieser fällt keine bedeutende ethische Funktion im Datenökosystem zu, aber sie ist notwendig, um ein Datenökosystem aus einer technischen Perspektive vollständig darzustellen.

Laut dem Betriebswirtschaftler Marco Iansiti und dem Technologieberater Roy Levien sowie Harrison et al. fällt zudem das Erstellen einer technischen Plattform zum Teilen der Daten in den Aufgabenbereich des keystone actors.[209] Aus meiner Sicht überschneidet sich dies mit den Aufgaben der sogenannten Serviceprovider, welche Oliveira et al. so beschreiben: „The service provider is responsible for producing and supporting software resources such as tools, applications, services, libraries, or other software products (...)."[210] Ein Service im Datenökosystem unterstützt die Analyse von Daten, um zur Wertschöpfung beizutragen.[211] Deshalb ist es die Aufgabe der Serviceprovider, zum einen die Infrastruktur für das Datenöko-

205 Nardi, O'Day 1999:53
206 Ebd.
207 Oliveira et al. 2019:608, 609, 610
208 van Schalkwyk et al. 2016:77
209 Iansiti, Levien 2004; Harrison et al. 2012:911
210 Oliveira et al. 2019:608
211 Immonen et al. 2014:89

system bereitzustellen und zum anderen die Analyseergebnisse für die Datennutzenden zu visualisieren. Die Bereitstellung der Infrastruktur ermöglicht die Datenspeicherung sowie das Erstellen und Verwalten der digitalen Plattform, auf der die verarbeiteten Daten für Datennutzende zur Verfügung gestellt werden können.[212] Die Wissenschaftler Juho Lindman, Tomi Kinnari und Matti Rossi aus der angewandten Informatik sind der Auffassung, dass die Aufgaben eines Serviceproviders von Unternehmen übernommen werden und zudem die Beratung von Kunden bezüglich ihrer Möglichkeiten zur Nutzung der Daten umfassen.[213] Für ein erfolgreiches Datenökosystem müssten Serviceprovider zudem die Informationen aktuell und zugänglich halten und dafür technische Services anbieten.[214] Im Folgenden nenne ich diejenigen, die die Analyse von Daten durch technische Services unterstützen, Infrastrukturbereitstellende.

Immonen et al. schreiben, dass Serviceprovider aus den Daten Informationen und Wissen produzieren.[215] Das ist aber nicht ihre Aufgabe, denn ein Serviceprovider unterstützt zwar die Analyse der Daten durch das Verwalten der Infrastruktur, aber er analysiert nicht selbst. Dafür gibt es die separate Rolle der Analysten, deren Funktion in der Literatur oft auch als Service beschrieben wird. Die Informatiker:innen Ying Chen, Jeffrey Kreulen, Murray Campbell und Carl Abrams zum Beispiel führen im Jahr 2011 Analytics as a Service (AaaS) provider als neue Entität im Datenökosystem ein.[216] Wer eine Analyse als Service bereitstellt, bietet technische Werkzeuge sowie die Durchführung der Analyse selbst an.[217] Ein solcher Service kann laut Immonen et al. aus Eingabedaten relevante Daten für einen bestimmten Kontext oder Bereich erzeugen. Diese kontextspezifischen Daten, die durch die Analyse entstehen, sollen bedeutsame Informationen sein – dadurch entstehe die Wertschöpfung im Datenökosystem.[218] Die Bereitstellenden von Analysen als Service bieten zwar im weitesten Sinne einen Service an, indem ihre

212 Ha et al. 2014:166, 168, 169; Lindman et al. 2016:5
213 Lindman et al. 2016:5
214 Ebd.:9
215 Immonen et al. 2014:95
216 Chen et al. 2011:14
217 Ebd.:12, 15, 17
218 Immonen et al. 2014:89

Datenanalyse eine Dienstleistung sein kann, jedoch ist ihre Rolle im Datenökosystem die der Datenverarbeitenden.

Chen et al. haben außerdem den Data as a Service (DaaS) provider als Akteur im Datenökosystem identifiziert. Sie sammeln, generieren und aggregieren Daten, bieten aber auch Datenzugriff, -verarbeitung und -verwaltungsdienste an.[219] Wer Daten als Service anbietet, sammelt und verwaltet viele verschiedene Datenquellen, sodass rohe Daten zur Verarbeitung vorbereitet und den Analysten angeboten werden können.[220] Solange sich die Arbeit mit den Daten auf die Vereinheitlichung des Datensatzes zur Vorbereitung für Datenverarbeitende beschränkt, fällt dies unter die Aufgaben der der Datengenerierenden.

Daten zur Nutzung frei geben, Daten generieren und Daten nutzen können viele Menschen – Daten analysieren können nur wenige. Deshalb ist die Funktionsfähigkeit eines Datenökosystems überaus abhängig von denen, die die Daten analysieren können. Darum bilden die Datenverarbeitenden das Herzstück eines Datenökosystems. Allerdings kann ihr Zugang zum Datenökosystem und auch dessen Sicherheit gefährdet werden, wenn die Infrastruktur von Dienstanbietenden außerhalb der rechtlichen oder territorialen Zuständigkeit der Datenverarbeitenden erstellt und verwaltet wird.[221] Deshalb ist es ratsam, dass die Verantwortung für die Infrastruktur und die Datenverarbeitung in einer Hand liegt, indem sie zum Beispiel von Mitarbeitenden des gleichen Unternehmens übernommen wird.

Die Verwaltung, die Verbindungen zwischen den Akteuren herstellt, könnte man im Datenökosystem da vermuten, wo am meisten von der Vermittlung profitiert wird. Es ist durchaus denkbar, dass die Datenquellen auch die Datennutzenden sein wollen und also ein großes Interesse an der Verarbeitung ihrer Daten haben. Es kann aber auch sein, dass die Analysten aus dem Team der Datenverarbeitenden einen Algorithmus erstellt haben, der viele Daten benötigt, um zu lernen und sich zu entwickeln. In dem Fall liegt den Datenverarbeitenden viel an der Vermittlung zwischen den Akteuren im Datenökosystem, da sie zum Training ihres Algorithmus darauf angewiesen sind, dass die Datengenerierenden ihnen

219 Chen et al. 2011:14
220 Ebd.:12, 14, 17
221 Shin 2016:842

Daten der Datenquellen zuspielen. Die Funktion der Verwaltung als Übersetzende für die effektive Nutzung von Technik[222] ist ebenfalls sehr nützlich für die Datenverarbeitenden: Jemand, der die Daten versteht, muss erklären, welches Potenzial der Datensatz hat und jemand, der weiß, was Datennutzende wollen, muss das Ziel in einen technischen Auftrag umwandeln können. Programmierende müssen verstehen, welche Fragen beantwortet werden sollen, und diejenigen, die verarbeitete Daten für Datennutzende bereitstellen, müssen die Antworten des Algorithmus verstehen. Die Vermittlung zwischen Datenquellen und Datenverarbeitenden übernehmen hingegen die Datengenerierenden. Wenn es also um die Vermittlung von Daten geht, liegt diese Tätigkeit bei den Datengenerierenden.
Wenn man aber betrachtet, wer Verantwortung tragen sollte für Regeln, Überwachung und Maßnahmen im Datenökosystem[223], wird deutlich, dass die Verwaltenden im Idealfall unabhängig von den Akteuren agieren, sodass niemand bevorzugt oder benachteiligt wird. Aber welcher der Akteure sorgt dafür, dass das Datenökosystem verwaltet wird? Ich behaupte, es ist der Akteur, der die Entstehung des Datenökosystems zu einem bestimmten Zweck anstößt.

Da bisher keins der vorgestellten Modelle aus der Literatur alle der von mir aufgestellten Kriterien erfüllt, erstelle ich an dieser Stelle ein neues Modell der Struktur eines Datenökosystems. Das heißt, es werden die Objekte dargestellt, welche das System ausmachen und es wird die Beziehung zwischen den Objekten hervorgehoben.[224] Dabei werden außerdem alle Elemente eines Datenökosystems dargestellt.[225] Denn die Akteure sind die Objekte des Systems. Nicht alle der bisher besprochenen Akteure, wie zum Beispiel die Verwaltung, haben ethische Relevanz und damit Relevanz für diese Arbeit.[226] Deshalb werde ich mich im Folgenden auf die vier Akteure mit ethischer Relevanz konzentrieren: Personen, die ihre Daten zur Verfügung stellen (*Datenquellen*), verschiedene Beteiligte, die die Daten generieren (*Datengenerierende*), diejenigen, die die Daten kombinieren und analysieren, um sie dann aufbereitet zur Verfügung stellen

222 Nardi, O'Day 1999:53
223 Oliveira et al. 2019:608, 609, 610
224 Shapiro 1997:74
225 Adner 2017:43f.; Oliveira, Lóscio 2018:8
226 Was ethische Relevanz hier bedeutet wird im Laufe der Arbeit deutlich.

zu können (*Datenverarbeitende*) und die Personen, die mit den aufbereiteten Daten arbeiten (*Datennutzende*).

Besagte *Datenquellen* sind all die Akteure, die Daten erzeugen. Diese Daten können zum Beispiel durch Messungen erzeugt werden, aber jede Person ist auch an sich eine Datenquelle. Denn Daten, die sich auf die eigene Person beziehen, erzeugen wir Menschen allein durch unsere Existenz. So sind zum Beispiel Name und Wohnort Daten, mit denen eine Person identifiziert werden kann, aber auch das Verhalten im Internet erzeugt Daten, die persönliche Präferenzen abbilden können. Deshalb kann prinzipiell jede Person eine Datenquelle sein. Der Datenfluss im Datenökosystem beginnt aber erst dann, wenn die Daten für *Datengenerierende* von Interesse sind. Denn Datengenerierende sammeln und speichern die Daten, meist schon zum Zweck der Verarbeitung. Entweder sie sind dabei initiativ und suchen selber nach für ihr Projekt passenden Daten, oder sie bieten gemeinsam mit den Datenverarbeitenden gewisse Ergebnisse an, wenn Datenquellen bestimmte Daten über sich preisgeben: Sie sollen zum Beispiel ihre Schulnoten angeben und bekommen eine Berufsempfehlung, oder sie sollen unter anderem Einkommen und Wohnungsgröße angeben und erfahren, ob sie Anspruch auf Wohngeld haben. Damit diese versprochenen Ergebnisse zustande kommen, braucht man die *Datenverarbeitenden*. Sie kombinieren und analysieren die von den Datengenerierenden gesammelten Daten mit Hilfe von Algorithmen und bereiten die algorithmischen Ergebnisse auf, um sie den Datennutzenden zur Verfügung zu stellen. Welche Aussagekraft die Ergebnisse haben, hängt dabei, wie oben erläutert, von den analysierten Daten ab. Die *Datennutzenden* werden mit den Ergebnissen konfrontiert. Wenn sie auch schon bewusst als Datenquelle gedient haben, ist ihnen klar, dass die Ergebnisse eine Aussage über sie machen und ihnen gewissen Möglichkeiten bieten oder verwehren können. Wenn die Daten der Datenquellen aber ohne ihr Zutun gesammelt und verarbeitet wurden, wie zum Beispiel durch Cookies bei der Internetbenutzung, dann ist es auch möglich die Datenquellen durch personalisierte Werbung zu beeinflussen – Datennutzende sind in dem Fall dann Werbende, welche die Ergebnisse nutzen, um ihre Produkte bei den Datenquellen anzupreisen, denen sie gefallen könnten.

Hier wird deutlich, dass Daten die Ressourcen des Datenökosystems darstellen. Außerdem wird durch die Beschreibung erkennbar,

dass jeder Akteur eine gewisse Rolle im System hat, sodass jedem Akteur bestimmte Aktivitäten zugeschrieben werden können. So entstehen im Fluss der Aktivitäten gewissen Positionen im System, da manche Akteure mehr Kontrolle über das Vorgehen im Datenökosystem haben als andere. All diese Elemente bedingen die Beziehungen zwischen den Akteuren.[227]

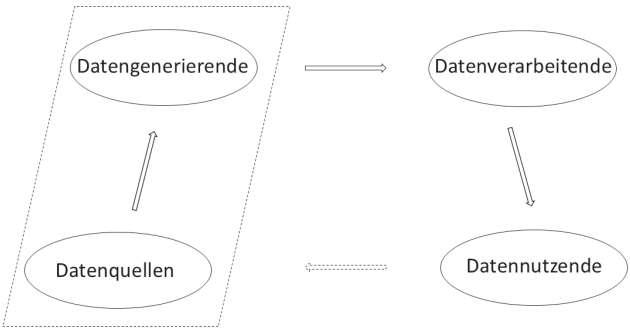

Abbildung 3 Ethisch relevante Akteure eines Datenökosystems[228]

Mein Modell der ethisch relevanten Akteure eines Datenökosystems in Abbildung 3 zeigt, dass die Daten von den Datenquellen zu den Datengenerierenden fließen, von da aus zu den Datenverarbeitenden und nach der Analyse gelangen sie weiter zu den Datennutzenden. Dabei können Datenquellen und Datengenerierende auch ein Akteur sein, was durch den gestrichelten Rahmen in Abbildung 3 dargestellt wird. Wenn zum Beispiel Hersteller von Flugzeugteilen Wartungsdaten erzeugen, sammeln und weitergeben, dann sind diese Hersteller gleichzeitig Datenquellen und Datengenerierende.[229] Dieses Detail ist wichtig, weil nur bei separaten Datenquellen und Datengenerierenden ein ungleiches Kontrollverhältnis zwischen diesen Akteuren bestehen kann und nur dann an dieser Stelle eine ethische Gefahr lauert. Auch Datengenerierende und Datenverarbei-

227 Adner 2017:43f.; Oliveira, Lóscio 2018:8
228 Eigene Darstellung
229 Aviatar 2022

tende können dem gleichen Unternehmen angehören, aber da es sich um zwei verschiedene Rollen handelt, sind diese Akteure hier separat dargestellt. Die Datennutzenden können gleichzeitig auch die Datenquellen gewesen sein – in diesem Fall schließt sich der Kreis der Datennutzung. Diese Akteure sind im Modell jedoch ebenfalls separat dargestellt, da auch sie unterschiedliche Rollen im Datenökosystem spielen. Somit erfüllt mein Modell alle drei zuvor aufgestellten Kriterien. Denn es sind sowohl der Datenfluss als auch die beteiligten Akteure und ihre Rollen klar erkennbar.

Die Pfeile in Abbildung 3 verdeutlichen den Datenstrom und zeigen auf, welcher Akteur die Daten an welchen anderen Akteur weiterreicht. Das Weiterreichen von Datenverarbeitenden zu Datennutzenden beinhaltet dabei das Anbieten der aufbereiteten Daten über eine digitale Plattform, die auch die Vermittlung zwischen den Datennutzenden und den Datenquellen ermöglicht. Diese Aufgabe des Anbietens der Daten über eine digitale Plattform könnte als eigenes Element bezeichnet werden, da sie ein bedeutender Schritt in der Datenvermittlung ist.[230] In meinem Modell wird das Anbieten der Daten jedoch mit dem Pfeil von Datenverarbeitenden zu Datennutzenden dargestellt und somit als Teil des Datenflusses definiert. Denn das reine Anbieten der Daten hat keine ethischen Implikationen, solange erklärt wird, wie die angebotenen Ergebnisse zustande gekommen sind. Diese Erklärung erwarte ich von den Datenverarbeitenden. Sie muss über die datenanbietende Plattform zusammen mit den Daten weitergegeben werden.

Anhand des Beispiels der Corona-Pandemie lässt sich verdeutlichen, welche Rolle Datenökosysteme für staatliches Handeln in Krisen spielen könnten. Patient:innen mit einer Corona-Erkrankung könnten Daten über ihren Krankheitsverlauf und die Wirksamkeit verschiedener Behandlungen an Ärzt:innen weitergeben – die Patient:innen würden dadurch zu Datenquellen. Die Ärzt:innen würden die Daten in ihrer Rolle als Datengenerierende an Datenverarbeitende weiterleiten. Letztere könnten mit Hilfe von Algorithmen große Mengen von Patient:innendaten verarbeiten und Informationen generieren, die sowohl durch die Patient:innen selbst, als auch Ärzt:innen oder den Staat genutzt werden könnten. Dieses Vorgehen könnte zeigen, welche Behandlungsmethoden bei wem am besten

230 Shah et al. 2020:108f.

anschlagen und sogar Aufschluss geben über die Gründe für einen schweren Krankheitsverlauf. So unterstützen Datenökosysteme die Verarbeitung und Analyse von Daten und damit das Generieren von Informationen, die für staatliches Handeln in der Krise hilfreich sein können.

Im Idealfall erfüllen alle Akteure ihre Rolle im Datenökosystem, weil sie einen Anreiz dazu haben, zum Beispiel, weil sie selbst einen Vorteil daraus ziehen. Das heißt, die Akteure werden entweder monetär entlohnt oder profitieren von aussagekräftigen Informationen, die durch ihre Teilnahme entstehen.[231] Datenquellen hätten den größten Anreiz, ihre Daten zur Verfügung zu stellen, wenn sie dafür bezahlt würden oder wenn sie selbst die verarbeiteten Daten nutzen könnten. De facto wissen Datenquellen jedoch oft nicht, dass bzw. wofür ihre Daten zusätzlich verwendet werden und können somit nicht von ihrer Rolle im Datenökosystem profitieren. Informierte Datenquellen können aber bewusst zu Datennutzenden werden und dadurch den Datenkreislauf schließen. Dabei bleiben Datenquellen und Datennutzende trotzdem zwei eigenständige Akteure, da sie unterschiedliche Motivationen haben. Der Anreiz für Datengenerierende, die Daten zu sammeln, ist genau der Gleiche – sie profitieren von Bezahlung oder den ausgewerteten Daten. Wenn zum Beispiel eine Schulleitung die Leistungsdaten ihrer Schüler:innen sammelt, um sie analysieren zu lassen, will sie auf Grundlage der verarbeiteten Daten ihr Lehrangebot verbessern.

Es ist denkbar, dass Datenverarbeitende sogar bereit wären für diese Daten zu zahlen. Denn wenn sie ein KI-System entwickelt haben, muss der Algorithmus mit vielen Daten trainiert werden, damit er seine Ergebnisse optimieren kann. Somit haben Daten (von den Quellen selbst oder von Datengenerierenden) einen großen Wert für Datenverarbeitende. Schließlich können Datenverarbeitende die Ergebnisse ihres Algorithmus an Datennutzende verkaufen und somit ihre Investition in die Daten wieder hereinholen. Wenn Datenquellen oder Datengenerierende zu Datennutzenden werden, ist der Datenkreislauf vollständig und Daten sind die einzig notwendige Währung.

231 Oliveira et al. 2019:618