https://datascience-project.eu/

Training Fiche Vorlage

Titel	Data Science und gesellschaftliche Auswirkungen: Positive Ergebnisse erzielen
Schlüsselwörter (meta tag)	Gesellschaftliche Auswirkungen, Daten für den guten Zweck, Fairness-Metriken, Überwachung sozialer Medien
Sprache	Deutsch
Zielsetzungen / Lernziele / Lernergebnisse	1. **Einsatz von Datenwissenschaft für das Gemeinwohl** 2. **Die Hauptrisiken der Technologie verstehen und Beispiele nennen können** 3. **Die Merkmale einer "vertrauenswürdigen KI" aufzählen können** 4. **Die Herausforderungen bei der Messung von Fairness verstehen**

Lehrgang:	
Datenwissenschaftliche Kompetenz	
Modul Datenvisualisierung und visuelle Analyse	
Einführung in die Datenwissenschaft für Human- und Sozialwissenschaften	
Datenwissenschaft für den guten Zweck	X
Datenjournalismus und Geschichtenerzählen	

Beschreibung	In diesem Kurs werden wir einen Blick auf die vielen datenwissenschaftlichen Anwendungen werfen, die die Welt ein wenig besser machen können. Anschließend werden wir einen genaueren Blick auf die Überwachung Sozialer Medien am Beispiel von Amnesty International Italien werfen, um besser zu verstehen, wie solche Anwendungen funktionieren können. Im nächsten Abschnitt werden wir mögliche schädliche Auswirkungen untersuchen, die Data Science und KI mit sich bringen können. Dies wird uns helfen, zu verstehen, warum KI-Systeme vertrauenswürdig sein müssen. Schließlich werden wir uns mit einigen der Herausforderungen von Fairness-Metriken vertraut machen und sehen, was diese Metriken in der Praxis bedeuten können.
Inhalt in 3 Ebenen gegliedert	1. **Einsatz von Datenwissenschaft für das Gemeinwohl** Anhand verschiedener Anwendungsfälle, insbesondere des "Amnesty Italy Anwendungsfall", erhältst du einen Überblick, wie Data Science für gute Zwecke eingesetzt werden kann.

Co-funded by the Erasmus+ Programme of the European Union

Mit Unterstützung des Programms Erasmus+ der Europäischen Union. Dieses Dokument und sein Inhalt spiegeln die
Die Verantwortung für den Inhalt dieser Veröffentlichung trägt allein der Verfasser; die Kommission haftet nicht für die weitere Verwendung der darin enthaltenen Angaben.

1.1 Überblick über Data Science für gute Anwendungsfälle

Wie positiv sich die Datenwissenschaft auf die Menschen und den Planeten auswirken kann, lässt sich am besten an einigen Beispielen aus jüngerer Vergangenheit ablesen.

Der rasche technologische Wandel führt auch zu Veränderungen auf dem Arbeitsmarkt - alte Arbeitsplätze und Berufe verschwinden und werden durch neue ersetzt. Dies hat zur Folge, dass in einigen Sektoren Arbeitslosigkeit entsteht, während in anderen Sektoren Arbeitgeber:innen nur schwer qualifizierte Mitarbeiter:innen finden können. Viele der in den "verschwindenden" Sektoren erworbenen Qualifikationen könnten jedoch leicht angepasst und in neuen Sektoren wiederverwendet werden. Im Rahmen des Pilotprojekts SkillsFuture Singapore wird Datenwissenschaft eingesetzt, um solche "wiederverwendbaren" Fähigkeiten zu erkennen und Arbeitslosen mit gezielten Schulungen zu helfen, ihre Fähigkeiten an den Bedarf der expandierenden Branchen anzupassen.

KI kann auch eingesetzt werden, um die Vorhersagefähigkeit digitaler Zwillinge zu verbessern, zum Beispiel um die Lieferkette widerstandsfähiger zu machen. Digitale Zwillinge nutzen Daten, die einem Unternehmen zur Verfügung stehen - seien es Daten, die intern durch betriebliche, transaktionale oder andere Prozesse generiert werden, oder öffentlich verfügbare Daten wie die Wetterüberwachung - um die Lieferkette zu simulieren. KI-Systeme, die mit Reinforcement Learning (verstärkendes Lernen) trainiert wurden, können zu diesen digitalen Zwillingen hinzugefügt werden, so dass Unternehmen die Auswirkungen verschiedener "Was-wäre-wenn"-Szenarien, z. B. die Auswirkungen eines Tornados, erforschen und Maßnahmen zur Reaktion auf solche Szenarien entwickeln können [2].

KI-Systeme können auch auf vielfältige Weise eingesetzt werden, um Klimaziele zu erreichen. So setzt *Fero Labs* KI ein, um Stahlherstellern dabei zu helfen, die Verwendung von abgebauten Rohstoffen um bis zu 34% zu reduzieren und damit schätzungsweise 450.000 Tonnen CO2-Emissionen pro Jahr zu vermeiden. Das *Mapping the Andean Amazon Project* setzt KI zur Überwachung der Entwaldung über Satellitenbilder ein, um illegale Abholzung aufzudecken und politische Maßnahmen zu unterstützen [3].

Eine der Herausforderungen im Zusammenhang mit Elektrofahrzeugen besteht darin, dass sie Zugang zu einer speziell für sie konzipierten elektrischen Infrastruktur benötigen - nämlich zu Stromtankstellen. Wenn viele Autos gleichzeitig die gleiche Infrastruktur benötigen, kann dies eine erhebliche Herausforderung für das Stromnetz darstellen. Eines der Hindernisse für die großflächige Nutzung erneuerbarer Energiequellen sind die starken Schwankungen von Energieverfügbarkeit und begrenzte Kapazitäten zur

Co-funded by the Erasmus+ Programme of the European Union

Mit Unterstützung des Programms Erasmus+ der Europäischen Union. Dieses Dokument und sein Inhalt spiegeln die
Die Verantwortung für den Inhalt dieser Veröffentlichung trägt allein der Verfasser; die Kommission haftet nicht für die weitere Verwendung der darin enthaltenen Angaben.

Speicherung von Strom zu Zeiten der Spitzenverfügbarkeit, um ihn dann zu Spitzenverbrauchszeiten abzugeben. Fahrzeug-zu-Netz-Technologien, die es ermöglichen, Elektroautos als "Speicher" für überschüssige Energie zu nutzen und dem Netz Energie aus den Autos zu entziehen, wenn diese nicht in Gebrauch sind, können helfen, das Problem zu mildern. Mit Hilfe von KI hat *Caltech* ein adaptives Ladesystem entwickelt, das auf Grundlage der vom Fahrer angegebenen Abfahrtszeiten festlegt, wann welches Fahrzeug geladen wird und wann und wie viel Energie ins Netz zurückgespeist werden kann. Dies senkt die Gesamtbelastung des Stromnetzes und eröffnet eine interessante Möglichkeit, wie Elektroautos Stromnetze entlasten können [4].

Lieferketten sind unglaublich komplex, was eine Herausforderung für Rechtsvorschriften wie den *Uyghur Forced Labor Prevention Act* der USA darstellt, der darauf abzielt, höhere Sozial- oder Umweltstandards bei Produkten durchzusetzen. Der Altana-Atlas kombiniert geografische Informationen über Unternehmensstandorte und -einrichtungen mit Daten über die Eigentumsverhältnisse von Unternehmen, um die Handelsbeziehungen zwischen den einzelnen Sektoren darzustellen. Dies hilft Unternehmen, solche Gesetze besser einzuhalten und selbst Maßnahmen gegen Probleme wie Zwangsarbeit zu ergreifen [5].

Windkraftanlagen sind eine wichtige Quelle für erneuerbare Energie, doch ihre Leistung hängt von einem schwer zu kontrollierenden Faktor ab: Wind. Dies stellt eine Herausforderung für das Energienetz, aber auch für die Vertriebsabteilung von Windenergieanbietern dar. Energie, die besser vorhersehbar ist, kann höhere Preise erzielen. Um den Business Case von Windfarmen zu unterstützen, hat DeepMind ein neuronales Netz entwickelt, das auf der Grundlage von Wettervorhersagen und historischen Betriebsdaten trainiert wurde und die Leistung des Windparks 36 Stunden im Voraus vorhersagen kann, wodurch der Wert der erzeugten Energie um 20 % gesteigert werden kann [6].

1.2 Anwendungsfall Amnesty Italien

Soziale Medien sind ein wichtiger Teil des öffentlichen Raums und somit Schauplatz für Meinungsbildung. Um zu untersuchen, wie sich der politische Diskurs zu Menschenrechtsthemen entwickelt und wie sich dies auf benachteiligte Gruppen auswirkt, führt Anmesty International Italien jedes Jahr eine Beobachtung mit dem Namen "Hate Barometer" (Barometre dell'Odio) durch und setzt dabei datenwissenschaftliche Techniken ein.
Die Daten werden über öffentliche Facebook- und Twitter-APIs aus einer von Amnesty zur Verfügung gestellten Liste von öffentlichen Konten und Profilen gesammelt. Der Beobachtungszeitraum umfasst in der Regel zwischen vier und acht Wochen (2021 wurde der Beobachtungszeitraum auf 16 Wochen ausgedehnt). Während dieses Zeitraums werden die Kommentare der aktivsten Konten nach dem Zufallsprinzip ausgewählt, was eine Menge von 30.000 bis 100.000 Kommentaren ergibt, die von geschulten Freiwilligen von Amnesty

Co-funded by the Erasmus+ Programme of the European Union
Mit Unterstützung des Programms Erasmus+ der Europäischen Union. Dieses Dokument und sein Inhalt spiegeln die
Die Verantwortung für den Inhalt dieser Veröffentlichung trägt allein der Verfasser; die Kommission haftet nicht für die weitere Verwendung der darin enthaltenen Angaben.

hinsichtlich Thema und des Grad von Anstößigkeit gekennzeichnet werden. Alle Kennzeichnungen werden gegengeprüft, d. h., jeder Kommentar wird von zwei bis drei Freiwilligen gekennzeichnet, und etwaige Unstimmigkeiten werden vom Anstößigkeits-Gremium (Tavolo dell'Odio) von Amnesty geklärt.

Beispiel: Wahlen zum Europäischen Parlament 2019

In der Zeit vor den Wahlen zum Europäischen Parlament 2019 wurden in den sechs Wochen vor der Wahl (15. April - 24. Mai 2019) die öffentlichen Profile von 461 Kandidat:innen auf Twitter und Facebook erfasst. Insgesamt wurden zunächst 27.000 Beiträge und 4 Millionen Kommentare erfasst. In einem zweiten Schritt wurde die Größe des Datensatzes reduziert, um den Umfang der Social-Media-Aktivitäten der Profile für die Freiwilligen bewältigbar zu machen und gleichzeitig sicherzustellen, dass alle Parteien, alle Regionen und mindestens eine Frau und ein Mann pro Partei im Datensatz vertreten sind. Auf diese Weise umfasste der endgültige Datensatz die Social-Media-Aktivitäten von 77 Politiker:innen: 80 % der Beiträge wurden von 150 freiwilligen Amnesty-Mitarbeiter:innen gekennzeichnet, dazu kam eine Zufallsstichprobe von 100.000 Kommentaren.
Die Ergebnisse [8] zeigen, dass Hassreden nicht zufällig verteilt sind, sondern in Gruppen auftreten. Auch wenn die Gesamtprävalenz auf Social-Media-Plattformen auf etwa 1 % geschätzt wird, ist es wahrscheinlicher, dass Hassreden im Zusammenhang mit bestimmten Gruppen und Themen auftreten und zu bestimmten Zeiten ihren Höhepunkt erreichen. Hassreden sind beispielsweise wahrscheinlicher, wenn es in der Diskussion um Migration, Roma, religiöse Minderheiten oder Frauen geht.
Wenn die Daten genauer betrachtet werden, können bestimmte Muster entdeckt werden. Hassreden ziehen mehr Hassreden nach sich, da es wahrscheinlicher ist, dass sie weitere Interaktionen nach sich ziehen (wie Reaktionen, Teilen oder Kommentare). Hassreden werden auch genutzt, um Menschen aktiv von Social-Media-Plattformen auszuschließen: So wurde beispielsweise während der Beobachtungs-Kampagne 2020 beobachtet, wie zwei Frauen gezielt durch Hassreden angegriffen und drei von Social-Media-Plattformen verdrängt wurden [9].

2. Datenwissenschaft ist nicht immer gut

Leider können KI und Datenwissenschaft, wie jede andere Technologie auch, für schlechte Zwecke eingesetzt werden oder unbeabsichtigte Folgen haben. Im Gegensatz zu anderen Werkzeugen automatisiert KI jedoch Entscheidungen für uns und hat daher ein noch größeres Potenzial, Schaden anzurichten. Deshalb müssen wir uns auch bewusst sein, dass KI und Datenwissenschaft negative Auswirkungen auf Menschen, Gesellschaft und Umwelt haben können.

2.1 Wichtige bekannte Beispiele

Co-funded by the Erasmus+ Programme of the European Union

Mit Unterstützung des Programms Erasmus+ der Europäischen Union. Dieses Dokument und sein Inhalt spiegeln die
Die Verantwortung für den Inhalt dieser Veröffentlichung trägt allein der Verfasser; die Kommission haftet nicht für die weitere Verwendung der darin enthaltenen Angaben.

Datenwissenschaft soll uns basierend auf Daten dabei helfen, bessere Entscheidungen zu treffen, indem sie es ermöglicht, große Mengen oder sehr unterschiedliche Arten von Informationen zu verarbeiten. Wie wir bereits gesehen haben, kann Datenwissenschaft genutzt werden, um Prozesse zu überwachen oder zu verbessern, die dazu beitragen, die Welt besser zu machen. Jüngste Ereignisse haben jedoch gezeigt, dass wir den Ergebnissen von Algorithmen nicht blind vertrauen können, insbesondere wenn diese Ergebnisse ernsthafte negative Auswirkungen auf unser Leben haben können.

Bekannte Beispiele für solche negativen Auswirkungen finden sich bei KI-Anwendungen in den Bereichen Gesundheit, Arbeit und Umwelt:

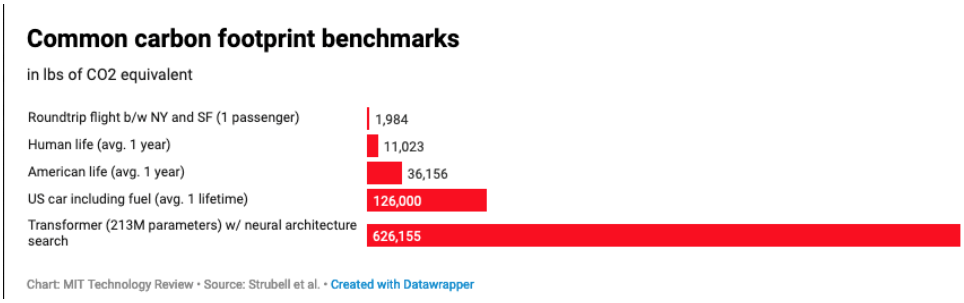
1. Krankenhäuser in den USA verlassen sich inzwischen auf Algorithmen, um zu beurteilen, wie krank Patient:innen sind und zu entscheiden, ob sie stationär oder ambulant behandelt werden müssen. In einer Studie wurde festgestellt, dass die Bewertungen eines sehr weit verbreiteten Systems rassistisch verzerrt waren: Schwarze Patient:innen waren in der Tat kränker als weiße Patient:innen, die die gleiche Risikoeinstufung erhalten hatten. Dies war wahrscheinlich darauf zurückzuführen, dass der Algorithmus die Gesundheitskosten der Vergangenheit als Indikator für den Gesundheitsbedarf heranzog - da das US-Gesundheitssystem jedoch seit jeher von Ungleichbehandlung geprägt ist, wurde weniger Geld für die Deckung des Gesundheitsbedarfs schwarzer Patient:innen ausgegeben. Der Algorithmus kam daher fälschlicherweise zu dem Schluss, dass sie gesünder sind als weiße Patient:innen, die in Wirklichkeit genauso krank sind [10].
2. Amazon entwickelte ein KI-Rekrutierungstool, um die Personalabteilung bei der Suche nach dem richtigen Personal für technische Stellen zu unterstützen, und trainierte es anhand von Lebensläufen, die dem Unternehmen in den letzten zehn Jahren vorgelegt wurden. Da die meisten dieser Bewerbungen jedoch von Männern stammten, stellte Amazon bald fest, dass sein Einstellungssystem die Kandidat:innen nicht geschlechtsneutral bewertete. Das KI-System bestrafte Lebensläufe, die von Frauen eingereicht wurden und Wörter wie "Frauen" enthielten. Die Software musste vom Netz genommen werden und wurde bis heute nicht wieder in Betrieb genommen [11].
3. Im Jahr 2015 bezeichnete Googles Bildklassifikator eine schwarze Person als "Gorilla". Google entschuldigte sich, entschied sich aber für eine schnelle Lösung, indem es einfach die Begriffe "Gorilla", "Schimpanse" und "Affe" aus der Suche und den Bild-Tags zensierte. Sechs Jahre später stufte Facebook einen schwarzen Mann in einem Video als Primaten ein und fragte

Co-funded by the Erasmus+ Programme of the European Union

Mit Unterstützung des Programms Erasmus+ der Europäischen Union. Dieses Dokument und sein Inhalt spiegeln die
Die Verantwortung für den Inhalt dieser Veröffentlichung trägt allein der Verfasser; die Kommission haftet nicht für die weitere Verwendung der darin enthaltenen Angaben.

Nutzer:innen, ob sie sich noch weitere Primatenvideos ansehen wollen. [12]

Dies sind nur einige Beispiele, die potenzielle negative Auswirkungen verdeutlichen. Datenwissenschaft und KI benötigen Daten - oft werden diese Daten von unterbezahlten Clickworker:innen etikettiert oder anderweitig verarbeitet, die unter sehr stressigen Bedingungen arbeiten und oft auch gewalttätigen oder verstörenden Inhalten ausgesetzt sind [13]. Algorithmen können verwendet werden, um Arbeitnehmer:innen oder Auftragnehmer:innen in einer Weise einzustufen, die diskriminierend ist und zu einem Verlust von Chancen führt [14]. Datenwissenschaft und KI sind rechenintensiv, was bedeutet, dass sie auch ressourcenintensiv sind. Dies gilt insbesondere für große Modelle und fein abgestimmte Modelle wie die Transformatoren in der Vergleichsgrafik unten [15].



Übung: Wenn du selbst zum Vorurteilsdetektiv werden willst, gehe einfach zu Google Translate (oder deepl.com) und übersetze vom Englischen ins Deutsche:

Englisch: My doctor is clever. She immediately found the solution.
Goolge Deutsch:

Englisch: My secretary is clever. He immediately found the solution.
Google Deutsch:

Nach einem großen Aufschrei in Bezug auf das Übersetzen von geschlechtsneutralen Sprachen in stereotype Geschlechterrollen hat Google 2018 versucht, dieses Problem anzugehen[1], aber wie du selbst feststellen kannst, gibt es auch fünf Jahre später noch Probleme.

2.2. Überblick über die wichtigsten Risiken

Anwendungen der Datenwissenschaft können Schaden verursachen: Von der Verwendung von Bots zur Erstellung von Deepfake-Nacktbildern auf Telegram

[1] https://blog.google/products/translate/reducing-gender-bias-google-translate/

Co-funded by the Erasmus+ Programme of the European Union

Mit Unterstützung des Programms Erasmus+ der Europäischen Union. Dieses Dokument und sein Inhalt spiegeln die
Die Verantwortung für den Inhalt dieser Veröffentlichung trägt allein der Verfasser; die Kommission haftet nicht für die weitere Verwendung der darin enthaltenen Angaben.

über die Erstellung sexualisierter Avatare von Frauen (aber nicht von Männern) bis hin zur Nichtentwicklung von Funktionen, die für eine bestimmte Personengruppe nützlich sind, oder zur Untergrabung der Geschlechtsidentität durch binäre Klassifizierung.

Eines der Hauptrisiken bei KI und Datenwissenschaft besteht darin, dass wir davon ausgehen, dass die Technologie selbst - wie jedes andere Werkzeug - frei von Beurteilungen und menschlichen Fehlern ist. In dieser Theorie scheinen wir jedoch zu vergessen, dass wir diejenigen sind, die diese Systeme schaffen, die die Algorithmen auswählen, die Daten selektieren und entscheiden, wie und für wen das System eingesetzt werden soll. Daher ist es von grundlegender Bedeutung zu verstehen, dass datenwissenschaftliche Anwendungen - selbst bei besten Absichten – nicht objektiv und nicht neutral sind.

Überlege, was deine Anwendung leisten kann, wofür sie verwendet wird, wer einbezogen/ausgeschlossen wird und wer auf unterschiedliche Weise betroffen sein könnte - die Folgen können weitreichend sein!

In ihrer Studie aus dem Jahr 2018 [15] fanden Joy Buolamwini und Timnit Gebru heraus, dass Algorithmen zur Geschlechterklassifizierung mittels Gesichtserkennung dunkelhäutige Frauen vergleichsweise häufiger falsch klassifizieren als hellhäutige Männer (und Frauen). Dies liegt daran, dass die Datensätze, auf denen die untersuchten Modelle trainiert wurden, einen überproportionalen Anteil von Bildern hellhäutiger Männer und Frauen enthielten.

Zwei Studien aus dem Jahr 2019 haben gezeigt, dass Algorithmen, die zur Erkennung von beleidigenden Äußerungen auf Online-Plattformen eingesetzt werden, die unter schwarzen US-Amerikanern verbreiteten Sprachmuster mit größerer Wahrscheinlichkeit als beleidigend einstufen. Die Datensätze, auf denen die Algorithmen basieren, weisen eine weit verbreitete Voreingenommenheit gegenüber afroamerikanischem Englisch auf [16]. Dies zeigt, wie wichtig die Kennzeichnung des Datensatzes ist: Wenn die Daten auf eine voreingenommene Weise gekennzeichnet werden, sind auch die Ergebnisse voreingenommen.

-> Wir müssen anerkennen, dass datenwissenschaftliche Anwendungen nicht perfekt sind und ihre Fehler nicht zufällig verteilt sind: Tatsächlich neigen diese Systeme dazu, insbesondere bei historisch marginalisierten oder gefährdeten demografischen Gruppen häufiger zu versagen.

Darüber hinaus können datenwissenschaftliche Anwendungen sehr datenintensiv sein. Dies bringt widerum Probleme mit sich:

- Datenschutz: KI-Modelle, die auf immer mehr Daten angewiesen sind, schaffen Anreize für die Sammlung von Daten in verschiedenen Bereichen. Das bedeutet, dass eine große Menge Daten über Menschen gesammelt werden, was erhebliche Auswirkungen auf die Privatsphäre

Co-funded by the Erasmus+ Programme of the European Union

Mit Unterstützung des Programms Erasmus+ der Europäischen Union. Dieses Dokument und sein Inhalt spiegeln die
Die Verantwortung für den Inhalt dieser Veröffentlichung trägt allein der Verfasser; die Kommission haftet nicht für die weitere Verwendung der darin enthaltenen Angaben.

hat. Während es beispielsweise aus Sicht des Verbrauchers oder der Verbraucherin manchmal praktisch sein kann, zu wissen, wo genau sich ein Paket gerade befindet und es aus Sicht eines Postdienstleisters oder einer Postdienstleisterin praktisch sein kann, über solche Daten zu verfügen, um die Routen zu optimieren, bedeutet die Verfolgung des Fahrzeugs, mit dem ein Paket zugestellt wird, auch die Verfolgung der Person, die das Fahrzeug fährt.

- Datenschutz: Viele der gesammelten Daten können die Identifizierung von Personen ermöglichen und gelten daher als personenbezogene Daten - wie das eben genannte Beispiel der Paketverfolgung. Solche Daten können nicht nur im Nachhinein missbraucht werden, sondern auch dazu dienen, sie einzuschränken. Deshalb sieht die Allgemeine Datenschutzverordnung der EU eine strenge Politik der Datenminimierung vor.
- Schlechte Datenqualität: Vielleicht hast du schon einmal von der Redewendung "Garbage in, garbage out" gehört, um zu beschreiben, wie schlechte Datenqualität zu schlechten Ergebnissen führen kann. Das bedeutet, dass ein Modell oder Ergebnisse nicht besser werden, wenn Sie einfach nur viele Daten beinhalten. Im Gegenteil: Ein großer Datensatz, der schlecht beschriftet, schlecht verarbeitet und voller irrelevanter Informationen ist, wird die Ergebnisse verschlechtern. Denke daran: Die meiste Zeit, die du in Data-Science- und KI-Projekte investierst, ist der Erstellung eines hochwertigen Datensatzes gewidmet, damit die Zuverlässigkeit und Reproduzierbarkeit gewährleistet ist. Es ist die Mühe wert!

Um den Risiken von Data Science und KI entgegenzuwirken, wurden bisher über 80 verschiedene Ethikrichtlinien entwickelt: Zu den bekanntesten gehören die von internationalen Organisationen wie der OECD, der UNESCO und UNICEF, aber auch von großen Technologieunternehmen wie Google und Microsoft.

Das Problem mit diesen Ethikstandards ist, dass sie weder rechtsverbindlich noch durchsetzbar sind: Es gibt keine Konsequenzen bei Nichteinhaltung. Ethikstandards helfen uns, die richtige Richtung einzuschlagen, und geben uns Hinweise darauf, was falsch und was richtig ist. Der freiwillige Charakter solcher Initiativen bedeutet jedoch, dass sie im Grunde ein "nice to have" sind, anstatt ein "must do".

3. Vertrauenswürdige KI

In diesem Abschnitt werden wir uns mit den Merkmalen der so genannten "vertrauenswürdigen KI" befassen und analysieren, woher der Begriff stammt und warum er wichtig ist. Wir werden uns auf das Thema der unerwünschten Voreingenommenheit (Bias) konzentrieren, die zu Diskriminierung führen kann.

Co-funded by the Erasmus+ Programme of the European Union

Mit Unterstützung des Programms Erasmus+ der Europäischen Union. Dieses Dokument und sein Inhalt spiegeln die
Die Verantwortung für den Inhalt dieser Veröffentlichung trägt allein der Verfasser; die Kommission haftet nicht für die weitere Verwendung der darin enthaltenen Angaben.

Außerdem gehen wir auf Möglichkeiten, wie man Fairness mit Hilfe einer Verwirrungsmatrix messen kann, ein.

3.1 Vertrauenswürdige KI

Auch die Europäische Union hat ihre eigenen Ethikstandards erstellt, die sogenannten "Ethics Guidelines for Trustworthy Artificial Intelligence" [17]. Ein Dokument, das von der Hochrangigen Expert:innengruppe für Künstliche Intelligenz (AI HLEG) erstellt wurde, einer unabhängigen Expert:innengruppe, die von der Europäischen Kommission im Juni 2018 als Teil der KI-Strategie der EU eingerichtet wurde.

Die EU-HLEG hat auf der Grundlage der EU-Charta der Grundrechte die folgenden Merkmale eines vertrauenswürdigen KI-Systems festgelegt: [2]

(1) Menschliches Handeln und Kontrolle: KI-Systeme sollten von Menschen in dem Maße verstanden werden, dass ihre Entscheidungen angefochten werden können, und Menschen sollten immer in der Lage sein, in KI-Systeme einzugreifen.

(2) Technische Robustheit und Sicherheit: KI-Systeme sollten in der Lage sein, eine Vielzahl von Situationen zu bewältigen, mit denen sie vernünftigerweise konfrontiert werden könnten, sowie böswillige Angriffe zu bewältigen, und sie sollten mit Blick auf die Sicherheit konzipiert werden.

(3) Datenschutz und Datenverwaltung: KI-Systeme sollten das Recht auf Privatsphäre nicht untergraben, die betroffenen Personen sollten die volle Kontrolle darüber haben, wie ihre Daten verwendet werden. Die Daten sollten nicht dazu verwendet werden, den betroffenen Personen zu schaden oder sie zu diskriminieren. Darüber hinaus muss ein geeignetes Data-Governance-System vorhanden sein, um sicherzustellen, dass der Datensatz von hoher Qualität ist und nicht für unrechtmäßige Zwecke verwendet werden kann.

(4) Transparenz: Die von KI-Systemen getroffenen Entscheidungen sollten für Menschen nachvollziehbar und erklärbar sein, und die Grenzen des KI-Systems sollten klar kommuniziert werden.

(5) Vielfalt, Nichtdiskriminierung und Fairness: Voreingenommene Datensätze verursachen Probleme, aber auch voreingenommene Modelle oder KI-Systeme, die unverhältnismäßige Auswirkungen auf bestimmte - und meist benachteiligte - Gruppen haben, sind schädlich. Aus diesem Grund sind eine vielfältige Vertretung und Beteiligung in allen Phasen des KI-Entwicklungszyklus der Schlüssel zur

[2] Die Charta der Grundrechte der Europäischen Union fasst die wichtigsten persönlichen Freiheiten und Rechte der EU-Bürger in einem rechtsverbindlichen Dokument zusammen. Siehe z.B. https://fra.europa.eu/en/eu-charter

Co-funded by the Erasmus+ Programme of the European Union

Mit Unterstützung des Programms Erasmus+ der Europäischen Union. Dieses Dokument und sein Inhalt spiegeln die
Die Verantwortung für den Inhalt dieser Veröffentlichung trägt allein der Verfasser; die Kommission haftet nicht für die weitere Verwendung der darin enthaltenen Angaben.

frühzeitigen Erkennung möglicher Schäden und zur Entwicklung geeigneter Präventions- und Abhilfemaßnahmen.

(6) Ökologisches und gesellschaftliches Wohlergehen: KI-Systeme haben echte Auswirkungen auf die Gesellschaft und die Umwelt, nicht nur auf den oder die Einzelne:n. Dies bedeutet, dass der Einsatz von KI-Systemen in einigen Bereichen gut überlegt sein sollte und dass alle KI-Systeme in ökologischer und sozialer Hinsicht nachhaltig gestaltet sein sollten.

(7) Rechenschaftspflicht: KI-Systeme sollten überprüfbar sein, und potenzielle negative Auswirkungen sowie Kompromisse sollten im Voraus identifiziert und angegangen werden, so dass im Falle eines Schadens die Möglichkeit einer wirksamen Wiedergutmachung besteht.

Der EU-Leitfaden geht einen Schritt weiter als einfache Ethik-Leitlinien, indem er die Grundsätze in der EU-Grundrechtecharta (einem rechtlichen Rahmen) verankert. Jedoch werden wir im nächsten Abschnitt am Beispiel von Fairness und Nichtdiskriminierung (Grundsatz 5) sehen, dass es vom Grundsatz bis zur Umsetzung noch ein weiter Weg ist.

3.2. Voreingenommenheit, Fairness, Nicht-Diskriminierung

Wir alle haben ein Menschenrecht darauf, fair behandelt zu werden. Aber was ist mit Fairness gemeint? Grundsätzlich ist Fairness ein subjektives Konzept, das von kulturellen Aspekten und vom individuellen Kontext abhängt. In dem Versuch, dieses heikle Thema zu umgehen, konzentrierte sich ein Großteil der Forschung stattdessen auf die Frage der Voreingenommenheit (Bias) in KI.

Im Kontext der Datenwissenschaft und des maschinellen Lernens im Allgemeinen kollidieren jedoch viele verschiedene Definitionen von Voreingenommenheit (umgangssprachlicher Gebrauch vs. Statistik vs. Deep Learning). Dies ist ein Problem, weil Menschen mit unterschiedlichem disziplinärem Hintergrund über Voreingenommenheit sprechen, aber eigentlich nicht dasselbe meinen. Im Kontext der vertrauenswürdigen KI verstehen wir unter Voreingenommenheit ein Vorurteil, das eine Gruppe gegenüber einer anderen bevorzugt.

Es gibt viele verschiedene Arten von Voreingenommenheit (Bias), wie z. B. gesellschaftliche Voreingenommenheit, Bestätigungsvoreingenommenheit, gruppeninterne Voreingenommenheit, Automatisierungsvoreingenommenheit, zeitliche Voreingenommenheit, Voreingenommenheit bei ausgelassenen Variablen, Stichprobenvoreingenommenheit, Repräsentationsvoreingenommenheit, Messvoreingenommenheit, Bewertungsvoreingenommenheit und viele andere.

All diese Verzerrungen (Bias) - in den Daten, im KI-System oder durch die Interaktion von Menschen mit Vorurteilen mit dem KI-System - können zu ungerechter Behandlung und Diskriminierung führen. Dies beinhaltet ungerechte

Co-funded by the Erasmus+ Programme of the European Union

Mit Unterstützung des Programms Erasmus+ der Europäischen Union. Dieses Dokument und sein Inhalt spiegeln die
Die Verantwortung für den Inhalt dieser Veröffentlichung trägt allein der Verfasser; die Kommission haftet nicht für die weitere Verwendung der darin enthaltenen Angaben.

oder vorurteilsbehaftete Behandlung verschiedener Personengruppen, z. B. aufgrund von ethnischer Zugehörigkeit, Alter, Geschlecht oder Behinderung.

Doch wie lassen sich Verzerrungen erkennen und messen?

Der erste Schritt besteht darin, die Qualität der Daten zu überprüfen, denn dies ist eine der häufigsten Möglichkeiten, wie sich Verzerrungen in den Datensatz einschleichen können. Aber selbst wenn die Daten keine Mängel aufweisen, kann das Modell dennoch verzerrt sein.

Eine Verzerrung lässt sich in der Regel erst als Auswirkung auf das Ergebnis des Modells feststellen. Dies geschieht mit einer so genannten Fairness-Metrik, die Thema des nächsten Abschnitts ist. Wie du siehst, ist der Versuch, die Definition von Fairness zu vermeiden und stattdessen auf Verzerrungen zu achten, nicht sehr weit gediehen.

3.3. Fairness-Metrik

Da es keine einheitliche, perfekte Definition von Fairness gibt, gibt es auch nicht die eine richtige Messgröße, um Fairness zu messen und somit ist eine allumfassende Einheitslösung unmöglich. Stattdessen gibt es viele verschiedene Arten von Fairness und Möglichkeiten, sie zu messen, darunter Gruppenfairness, bedingte statistische Parität, Gleichgewicht der Falsch-Positiv-Fehlerquote, Gleichgewicht der Falsch-Negativ-Fehlerquote, Gleichheit der bedingten Verwendungsgenauigkeit, Gleichheit der Gesamtgenauigkeit, Test-Fairness, Wohl-Kalibrierung, Fairness durch Unwissenheit, kontrafaktische Fairness und viele mehr.

Leider kannst du nicht einfach alle testen, um sicherzustellen, dass dein Algorithmus fair ist, da diese Messgrößen wahrscheinlich zu widersprüchlichen Ergebnissen führen werden. So ist es zum Beispiel mathematisch unmöglich, sowohl die Anforderungen an die *Predictive Parity* als auch an die *Qqualized Odds* zu erfüllen. Betrachte die folgende, in [18] abgeleitete Formel:

$$FPR = \frac{p}{1-p}(1 - FNR)\frac{.1 - PPV}{PPV}$$

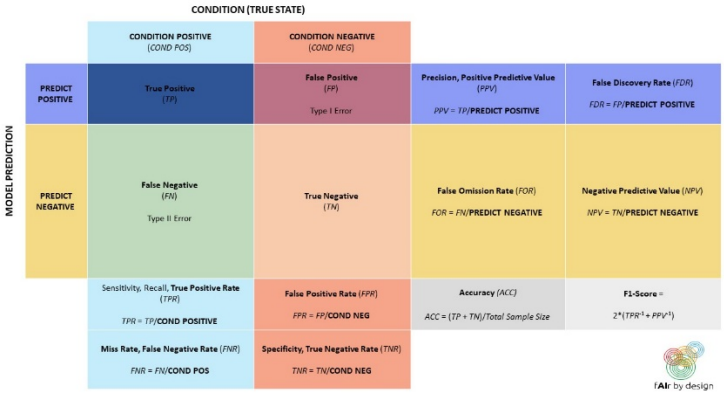
Das p in der Formel bezieht sich auf die Prävalenz der POSITIVEN Klasse und du kannst die untenstehende Vertauschungsmatrix verwenden, um die anderen Begriffe zu verstehen. Nehmen wir nun an, dass es zwei demografische Gruppen gibt, G1 und G2, mit den Prävalenzen p1 und p2. Wenn *Equalized Odds* gilt, sind FPR und FNR für beide Gruppen gleich. Wenn die *Predictive Parity* gilt, dann ist auch der PPV für beide Gruppen gleich. Werden all diese Informationen in die obige Formel eingesetzt, ergeben sich zwei Gleichungen, eine für G1 und eine für

Co-funded by the Erasmus+ Programme of the European Union

Mit Unterstützung des Programms Erasmus+ der Europäischen Union. Dieses Dokument und sein Inhalt spiegeln die
Die Verantwortung für den Inhalt dieser Veröffentlichung trägt allein der Verfasser; die Kommission haftet nicht für die weitere Verwendung der darin enthaltenen Angaben.

G2. Ein wenig Algebra wird dir dann zeigen, dass p1 und p2 ebenfalls gleich sein **müssen**.

Zusammenfassend lässt sich sagen: Wenn sowohl die *Equalized Odds* als auch die *Predictive Parity* zutreffen, dann muss die Prävalenz in beiden Gruppen gleich sein. Umgekehrt, wenn die Prävalenz nicht für beide Gruppen gleich ist, dann **können** *Equalized Odds* und *Predictive Parity* **nicht** beide zutreffen!



Die mathematische Unmöglichkeit, alle Fairnesskriterien gleichzeitig zu erfüllen, bedeutet, dass eine Entscheidung darüber getroffen werden muss, welche Definition von Fairness angewendet werden soll. Leider gibt es derzeit weder einen Rechtsrahmen noch Best-Practice-Beispiele. Das bedeutet, dass du den Kontext deiner KI-Anwendung sorgfältig abwägen musst, bevor du eine Metrik für die Bewertung ihrer Auswirkungen im Hinblick auf Fairness auswählst.

Um zu verstehen, welche Auswirkungen mehrere Definitionen von Fairness haben, die nicht miteinander vereinbar sind, und wie wichtig es ist, sich auf eine Definition zu einigen, bevor solche Systeme eingesetzt werden, werfen wir einen Blick auf ein reales Beispiel, das einen Großteil der Forschung und Debatte über Verzerrungen in Algorithmen in der Data Science- und ML-Gemeinschaft ausgelöst hat.

COMPAS ist ein KI-System, das von einem Unternehmen namens Northpointe entwickelt wurde und in der Strafjustiz der Vereinigten Staaten eingesetzt wird, um das Rückfallrisiko von Angeklagten einzuschätzen (mit anderen Worten: das Risiko eines oder einer Angeklagten, in Zukunft eine weitere Straftat zu begehen). Dieser Risikowert wird dann verwendet, um Entscheidungen über eine Bewährung oder eine vorzeitige Entlassung zu treffen.

Für die resultierenden Ergebnisse, griff das KI-System auf historische Verbrechensaufzeichnungen zurück, die frühere Straftäter:innen verfolgten und Auskunft darüber gaben, ob sie nach ihrer Entlassung wegen einer anderen Straftat erneut verhaftet wurden - D. h., das KI-System erhielt Informationen darüber, ob bestimmte Gruppen von Angeklagten wahrscheinlich erneut

Co-funded by the Erasmus+ Programme of the European Union

Mit Unterstützung des Programms Erasmus+ der Europäischen Union. Dieses Dokument und sein Inhalt spiegeln die
Die Verantwortung für den Inhalt dieser Veröffentlichung trägt allein der Verfasser; die Kommission haftet nicht für die weitere Verwendung der darin enthaltenen Angaben.

Straftätig werden (und dabei erwischt werden!). Zum Zeitpunkt der Inbetriebnahme, wurde das Modell anhand dieser Datensätze zur Vorhersage des Rückfallrisikos von neu Beschuldigten, die nicht Teil des Datensatzes waren, trainiert. Das bedeutet, dass die Rückfallwahrscheinlichkeit für jede:n Angeklagte:n berechnet wurde und die Angeklagten dann als risikoarm oder risikoreich eingestuft wurden.

Im Mai 2016 veröffentlichte ProPublica einen Artikel, der darauf hinwies, dass die Vorhersagen dieses Modells für Rückfälligkeit verzerrt waren [18; siehe auch 19, 20]: ProPublica wies nach, dass die Formel des KI-Systems schwarze Angeklagte besonders häufig fälschlicherweise als Personen mit hohem Rückfallrisiko einstufte, und zwar fast doppelt so häufig wie weiße Angeklagte (42 % gegenüber 22 %); gleichzeitig wurden weiße Angeklagte häufiger fälschlicherweise als Personen mit geringem Risiko eingestuft als schwarze Angeklagte.[3]

Wenn wir uns die obige Vertauschungsmatrix ansehen, können wir herauslesen, dass ProPublica sagte, dass COMPAS unfair sei, weil FPR und FNR für schwarze Angeklagte nicht gleich sind wie für weiße Angeklagte. Es stellt sich heraus, dass dies die Fairness-Metrik *Equalized Odds* ist:

1. Equalized Odds

Equalized Odds bedeutet, dass innerhalb jeder echten Risikokategorie der Prozentsatz der falsch negativen Vorhersagen und der falsch positiven Vorhersagen für jede Bevölkerungsgruppe gleich ist. Die Frage konzentriert sich nicht mehr auf die Gesamtgenauigkeit des Modells, sondern schlüsselt die Arten von Fehlern auf, die das Modell machen kann (falsch positive und falsch negative Vorhersagen), und verlangt, dass die Modellfehler gleichermaßen verteilt sind: Die FPR ist über alle Gruppen hinweg gleich, und die FNR ist über alle Gruppen hinweg gleich.

Northpointe verteidigte sein System COMPAS gegen den Vorwurf der Voreingenommenheit, indem es darauf hinwies, dass, wenn ein:e Angeklagte:r vom Modell als hochriskant vorhergesagt wurde, die Wahrscheinlichkeit, dass er oder sie tatsächlich erneut straffällig wird, gleich hoch ist, unabhängig davon, welcher demografischen Gruppe der oder die Angeklagte angehört. Northpointe will damit sagen, dass die Wahrscheinlichkeit eines tatsächlich positiven Ergebnisses, wenn das Modell ein positives Ergebnis vorhersagt, für alle Gruppen gleich hoch ist. Dies wird als "Predictive Parity"-Fairness-Maßstab bezeichnet.

2. Predictive Parity

Predictive Parity bedeutet, dass der Anteil der korrekt vorhergesagten Hochrisiko-Beschuldigten unabhängig von der demografischen Zusammensetzung gleich ist. Mit anderen Worten, Predictive Parity bezieht sich auf das Konzept in ML und KI,

[3] https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm

Co-funded by the Erasmus+ Programme of the European Union

Mit Unterstützung des Programms Erasmus+ der Europäischen Union. Dieses Dokument und sein Inhalt spiegeln die
Die Verantwortung für den Inhalt dieser Veröffentlichung trägt allein der Verfasser; die Kommission haftet nicht für die weitere Verwendung der darin enthaltenen Angaben.

dass die verwendeten Vorhersagemodelle denselben positiven Vorhersagewert (PPV) für verschiedene Gruppen erzeugen sollten, unabhängig von ihrer Zugehörigkeit zu einer schützenswerten Gruppe (z. B. Rasse, Geschlecht, Alter usw.). Der PPV ist eine Metrik zur Bewertung des Anteils wahrer positiver Vorhersagen (korrekt klassifizierte positive Instanzen) unter allen Instanzen, in denen das Modell eine positive Vorhersage bestimmte. Eine solche Kennzahl berücksichtigt jedoch nicht die Gesamtprävalenz von Instanzen innerhalb eines Datensatzes!

Um es noch einmal zu sagen: Predictive Parity berücksichtigt die Fairness, indem es die Fehler relativ zur *vorhergesagten* Klasse betrachtet, während Equalized Odds die Fehler relativ zur *wahren* Klasse betrachtet. Ob es wichtiger ist, das PPV zu optimieren (und du daher die Predictive Parity Fairness bevorzugen würdest), oder ob du lieber die FPR minimieren (und daher die Equalized Odds bevorzugst), ist eine Frage der Perspektive. Welche Fehlermetrik ist für dich beispielsweise wichtiger, wenn du eine medizinische Diagnose von einem KI-System erhalten hast? Und welche Fehlermetrik ist wichtiger für einen Einstellungsalgorithmus, der eine Stelle auswählt, auf die du dich beworben hast? Kannst du dir Situationen vorstellen, in denen du die PPV für wichtiger erachtest, und andere Situationen, in denen du eine niedrige FPR bevorzugen würdest?

Wenn du mehr über die verschiedenen Definitionen von Fairness (derzeit gibt es mehr als 21), ihre Messung und die Unterschiede zwischen ihnen erfahren möchtest, lies "Fairness Definitions Explained" [22].

Überlege dir: Um auf das COMPAS-Beispiel zurückzukommen, welche Definition würdest du als fair bezeichnen?

Erläutere: Ist es möglich, beide Definitionen (Equalized Odds und Predictive Parity) von fair zu erfüllen?

Antwort: Wir müssen die Prävalenz der Rückfälligkeit verstehen. In den USA ist die Rückfallquote für schwarze Angeklagte insgesamt höher als für weiße Angeklagte (52 % gegenüber 39 %). Nach der Formel, die wir oben gesehen haben, bedeutet dies, dass es nicht möglich ist, dass beide Definitionen von Fairness zutreffen.

Dieser COMPAS-Fall ist ein Beispiel dafür, wie sich soziale Fragen auf Daten auswirken die zur Verfügung stehen. Eine übermäßige Polizeipräsenz in schwarzen Gemeinden bedeutet, dass die Wahrscheinlichkeit von Verhaftungen oder registrierten Vorfällen in diesen Gemeinden höher ist. Infolgedessen fließen verzerrte Daten in die Modelle ein. Und noch subtiler - dies bedeutet, dass die wahrgenommene Rückfallquote für die beiden Bevölkerungsgruppen nicht dieselbe ist, was zu sehr schwierigen Entscheidungen darüber zwingt, welche Fairness-Metrik zu verwenden ist - d. h., was ist in diesem Zusammenhang überhaupt fair.

Co-funded by the Erasmus+ Programme of the European Union

Mit Unterstützung des Programms Erasmus+ der Europäischen Union. Dieses Dokument und sein Inhalt spiegeln die
Die Verantwortung für den Inhalt dieser Veröffentlichung trägt allein der Verfasser; die Kommission haftet nicht für die weitere Verwendung der darin enthaltenen Angaben.

Das eigentliche Problem besteht darin, dass es im Justiz- und Strafverfolgungssystem (in den USA, aber auch anderswo!) systembedingte Verzerrungen gibt, die nicht einfach dadurch behoben werden können, dass dem System mehr Daten (historische Fälle) zugeführt werden. Eine ausgezeichnete Diskussion über die Probleme bei der Verwendung schlechter Daten zur Erstellung von Prognosen in der Polizeiarbeit findet sich in "Dirty Data, Bad Predictions: How Civil Rights Violations Impact Police Data, Predictive Policing Systems, and Justice" [22].

Systembedingte Verzerrungen wirken sich auch auf andere Anwendungsbereiche aus, sei es im Gesundheitswesen, im Bildungswesen oder bei der Frage, wie viel du für Produkte oder Dienstleistungen bezahlst. Manchmal können wir die richtigen Instrumente wählen, um solche systemischen Verzerrungen zu berücksichtigen. Und manchmal müssen wir zugeben, dass die Bedingungen für eine sichere Verwendung von Algorithmen nicht gegeben sind. Solche Entscheidungen sollten jedoch nicht dem oder der Datenwissenschaftler:in allein überlassen werden, sondern eine Vielzahl von Interessengruppen und viele verschiedene Fachgebiete einbeziehen - darunter beispielsweise Soziologie, Psychologie, Recht und kontextspezifische Fachleute.

KI und Datenwissenschaft können keine Wunder vollbringen und unsere gesellschaftlichen Probleme lösen, aber wir können die Technologie als Instrument nutzen, um diese systemischen Probleme ans Licht zu bringen und sie als Gesellschaft insgesamt anzugehen.

Denn "KI funktioniert nur, wenn sie für uns alle funktioniert"[24].

4. Schlussfolgerung

Lass uns also zusammenfassen, was wir gelernt haben:

Einerseits gibt es für Datenwissenschaft und KI eine Vielzahl von Anwendungen mit positiven sozialen Auswirkungen. So ist die Datenwissenschaft beispielsweise nützlich, um zu untersuchen, wie sich soziale Medien auf die Menschenrechte auswirken. Andererseits bergen Datenwissenschaft und KI-Anwendungen auch Risiken für Gesundheit, Sicherheit, Umwelt und Menschenrechte. Voreingenommenheit (Bias) und Diskriminierung, Datenschutzbedenken und schädliche Auswirkungen auf die Umwelt sind nur einige der möglichen Folgen. Die Fairness der Ergebnisse von datenwissenschaftlichen und KI-Anwendungen kann auf viele verschiedene Arten gemessen werden. Die Entwicklung vertrauenswürdiger KI-Anwendungen erfordert eine intensive interdisziplinäre Zusammenarbeit: Wenn wir sicherstellen, dass unsere Entwicklungsprozesse integrativ sind und eine breite Beteiligung ermöglichen, können wir bessere Anwendungen entwickeln.

Co-funded by the Erasmus+ Programme of the European Union

Mit Unterstützung des Programms Erasmus+ der Europäischen Union. Dieses Dokument und sein Inhalt spiegeln die
Die Verantwortung für den Inhalt dieser Veröffentlichung trägt allein der Verfasser; die Kommission haftet nicht für die weitere Verwendung der darin enthaltenen Angaben.

Selbstbeurteilung (Multiple-Choice-Fragen und Antworten)	1. Nenne drei verschiedene Anwendungsfälle von Datenwissenschaft für den guten Zweck A) adaptives Aufladung B) Beurteilung von Fähigkeiten C) Überwachung sozialer Medien auf Auswirkungen auf die Menschenrechte 2. Welcher der folgenden Punkte gehört ***nicht*** zu den HLEG-Prinzipien der vertrauenswürdigen KI? A) Robustheit B) Reproduzierbarkeit C) Transparenz 3. Die Equalized Odds Fairness-Metrik erfordert, dass A) Die TPR für alle demografischen Gruppen gleich ist B) Die FPR für alle demografischen Gruppen gleich ist C) Alle der oben genannten Punkte
Ressourcen (Videos, Verweislinks)	- [1] Erkennung der Nachbarschaft von Fähigkeiten und gezieltes Training von fehlenden Fähigkeiten: SkillsFuture Singapur, https://www.skillsfuture.gov.sg/AboutSkillsFuture - [2] KI und digitale Zwillinge - Simulationen und Übungen für die Widerstandsfähigkeit der Lieferkette: https://www.technologyreview.com/2021/10/26/1038643/ai-reinforcement-learning-digital-twins-can-solve-supply-chain-shortages-and-save-christmas/ - [3] Verringerung des Fußabdrucks von recyceltem Stahl: Fero Labs setzt KI ein, um Stahlherstellern dabei zu helfen, die Verwendung von abgebauten Rohstoffen um bis zu 34 % zu reduzieren und so schätzungsweise 450.000 Tonnen CO2-Emissionen pro Jahr zu vermeiden: https://gpai.ai/projects/responsible-ai/environment/climate-change-and-ai.pdf - [4] Adaptives Laden beseitigt die Hindernisse für die Einführung von Elektrofahrzeugen. Bidirektionales Laden und Vehicle-to-Grid-Technologien erfordern intelligente Planungsalgorithmen. https://ev.caltech.edu/info - [5] Einsatz von KI zur Aufdeckung von Zwangsarbeit in der Lieferkette: https://www.altana.ai/blog/illuminating-xinjiang-forced-labor-ecosystem - [6] Maschinelles Lernen kann den Wert der Windenergie steigern: https://www.deepmind.com/blog/machine-learning-can-boost-the-value-of-wind-energy - [7] Barometer dell'Odio:. https://www.amnesty.it/campagne/contrasto-allhate-speech-online/ - [8] Barometre dell'Odio: Europäische Wahlen. https://d21zrvtkxtd6ae.cloudfront.net/public/uploads/2020/01/Amnesty-barometro-odio-2019.pdf - [9] Barometre dell'Odio: sessimo da tastiera. https://www.amnesty.it/barometro-dellodio-sessismo-da-tastiera/#sintesi - [10] Ziad Obermeyer et al. Dissecting racial bias in an algorithm used to manage the health of populations. https://science.sciencemag.org/content/366/6464/447 - [11] The Guardian, Amazon ditched AI recruiting tool that favored men for technical jobs, Oktober 2018. https://www.theguardian.com/technology/2018/oct/10/amazon-hiring-ai-gender-bias-recruiting-engine - [12] Nach den Gorillas von Google kommen die Primaten von Facebook: Facebook entschuldigt sich, nachdem die künstliche Intelligenz ein Video mit schwarzen Männern als "Primaten" gekennzeichnet hat, September 2021. https://www.nytimes.com/2021/09/03/technology/facebook-ai-race-primates.html - [13] - [14] - [15] Joy Buolamwini & Timnit Gebru. Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. http://proceedings.mlr.press/v81/buolamwini18a/buolamwini18a.pdf

Co-funded by the Erasmus+ Programme of the European Union

Mit Unterstützung des Programms Erasmus+ der Europäischen Union. Dieses Dokument und sein Inhalt spiegeln die
Die Verantwortung für den Inhalt dieser Veröffentlichung trägt allein der Verfasser; die Kommission haftet nicht für die weitere Verwendung der darin enthaltenen Angaben.

	- [16] Die Algorithmen, die Hassreden im Internet erkennen, sind gegen Schwarze voreingenommen. August 2019. https://www.vox.com/recode/2019/8/15/20806384/social-media-hate-speech-bias-black-african-american-facebook-twitter - [17] EU-HLEG-Leitlinien für vertrauenswürdige KI: https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai - [18] Chouldechova A. Fair Prediction with Disparate Impact: Eine Studie über Verzerrungen in Instrumenten zur Rückfallprognose. Big Data. 2017 Jun;5(2):153-163. - [19] Maschinelle Voreingenommenheit. Es gibt eine Software, die im ganzen Land eingesetzt wird, um zukünftige Verbrecher vorherzusagen. Und sie ist voreingenommen gegen Schwarze. Mai 2016. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing - [20] Ein Computerprogramm, das für Kautions- und Urteilsentscheidungen verwendet wird, wurde als voreingenommen gegen Schwarze bezeichnet. Das ist eigentlich nicht so klar. Oktober 2016. https://www.washingtonpost.com/news/monkey-cage/wp/2016/10/17/can-an-algorithm-be-racist-our-analysis-is-more-cautious-than-propublicas/ - [21] Julia Dressl und Hany Farid. Die Genauigkeit, Fairness und Grenzen der Rückfallprognose. Januar 2018. https://www.science.org/doi/10.1126/sciadv.aao5580 - [22] Sahil Verma, Julia Rubin: **"Fairness Definitions Explained",** 2018 ACM/IEEE International Workshop on Software Fairness; https://dl.acm.org/doi/10.1145/3194770.3194776 - [23] Richardson, R. et al, "Dirty Data, Bad Predictions: How Civil Rights Violations Impact Police Data, Predictive Policing Systems, and Justice"; https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3333423 - [24] D. Raji, "Wie unsere Daten systematischen Rassismus kodieren", MIT Technology Review. https://www.technologyreview.com/2020/12/10/1013617/racism-data-science-artificial-intelligence-ai-opinion/
Verwandtes Material	
Verwandte PPT	
Literaturverzeichnis	
Zur Verfügung gestellt von	[Women in AI Austria]

Co-funded by the Erasmus+ Programme of the European Union

Mit Unterstützung des Programms Erasmus+ der Europäischen Union. Dieses Dokument und sein Inhalt spiegeln die
Die Verantwortung für den Inhalt dieser Veröffentlichung trägt allein der Verfasser; die Kommission haftet nicht für die weitere Verwendung der darin enthaltenen Angaben.