Fairness und Künstliche Intelligenz: Warum Metriken nicht ausreichen

Originalartikel

Backup

<html> <header class=„article-header“><h1 class=„articleheading“>Fairness und Künstliche Intelligenz: Warum Metriken nicht ausreichen</h1> Isabel Bär </header><figure class=„aufmacherbild“><figcaption class=„akwa-caption“>(Bild: shuttersv/Shutterstock.com)</figcaption></figure> Klassisches Software-Testing lässt sich nicht ohne Weiteres auf KI übertragen. Model Governance und interne Audits sind nötig, um Fairness zu gewährleisten. Der Einsatz von Künstlicher Intelligenz (KI) bringt Verantwortung mit sich. Transparenz, Erklärbarkeit, Fairness sind dabei wesentliche Prinzipien, die ebenso gewährleistet sein müssen wie die hohe Leistungsfähigkeit des KI-Systems. Um diese Anforderungen einzuhalten, liegt es nahe, sich an Bereichen mit einer Tradition überprüfbarer Prozesse zu orientieren. Zwar funktionieren diese Prozesse nicht fehlerlos, aber ohne sie lassen sich Sicherheitsstandards nicht verwirklichen. Am offensichtlichsten ist das in sicherheitskritischen und regulierten Branchen wie der Medizin, aber auch in der Luft- und Raumfahrt oder im Finanzwesen. <header class=„a-boxheader“ data-collapse-trigger=„“>Young Professionals schreiben für Young Professionals</header><div class=„a-boxtarget a-boxcontent a-inline-textboxcontent a-inline-textboxcontent–horizontal-layout“ data-collapse-target=„“><figure class=„a-inline-textboximage-container“></figure> Dieser Beitrag ist Teil einer Artikelserie, zu der heise Developer junge Entwickler:innen einlädt – um über aktuelle Trends, Entwicklungen und persönliche Erfahrungen zu informieren. Die Reihe „Young Professionals“ erscheint im monatlichen Rhythmus. Bist du selbst ein „Young Professional“ und willst einen (ersten) Artikel schreiben? Schicke deinen Vorschlag an die Redaktion: developer@heise.de. Wir stehen dir beim Schreiben zur Seite. <ul class=„a-inline-textboxlist“><li class=„a-inline-textboxitem“><a class=„a-inline-textboxtext“ href=„https://www.heise.de/hintergrund/Alpine-js-Das-Schweizer-Taschenmesser-fuer-dynamische-Weboberflaechen-6177628.html“ title=„Alpine.js: Das Schweizer Taschenmesser für dynamische

Weboberflächen“>Alpine.js: Das Schweizer Taschenmesser für dynamische Weboberflächen [1]
<li class=„a-inline-textboxitem“>Developer Experience: Glückliche Entwickler schreiben besseren Code [2]
<li class=„a-inline-textboxitem“>Ethik und Künstliche Intelligenz: Ein neuer Umgang mit KI-Systemen [3]
<li class=„a-inline-textboxitem“>Alle Beiträge der Serie finden sich in der Rubrik „Young Professionals“ [4] Ähnlich wie diese Bereiche Prozesse benötigen, um relevanten Anforderungen nachzukommen, benötigt ein Unternehmen, das KI-Systeme einsetzt, geregelte Abläufe, durch die es Zugriff auf Machine-Learning-Modelle (ML) kontrolliert, Richtlinien sowie gesetzliche Vorgaben umsetzt, die Interaktionen mit den Modellen und deren Ergebnissen verfolgt sowie festhält, auf welcher Grundlage ein Modell erzeugt wurde. Insgesamt werden diese Prozesse als Model Governance bezeichnet. Model-Governance-Prozesse sind von Beginn an in jede Phase des ML-Lebenszyklus zu implementieren (Design, Development und Operations). Zur konkreten technischen Integration von Model Governance in den ML-Lebenszyklus hat die Verfasserin sich andernorts ausführlicher geäußert. <h3 class=„subheading“ id=„nav_model0“>Model Governance: ein Muss im Regel- und Auflagenwald</h3> Model Governance ist nicht optional (siehe Kasten „Checkliste Model Governance“). Zum einen gibt es bereits bestehende Regularien, die Unternehmen in bestimmten Branchen erfüllen müssen. Am Beispiel des Finanzsektors lässt sich die Bedeutung von Model Governance gut illustrieren: Kreditvergabesysteme oder Zinsrisiko- und Preisbildungsmodelle für Derivate sind risikoreich und verlangen ein hohes Maß an Kontrolle und Transparenz. Laut einer Algorithmia-Studie zu den wichtigsten Trends im KI-Einsatz für 2021 [5] ist die Mehrzahl der Unternehmen an die Erfüllung rechtlicher Auflagen gebunden – 67 Prozent der Befragten müssen mehreren Vorschriften entsprechen. Lediglich 8 Prozent gaben an, keinen gesetzlichen Vorgaben zu unterliegen. Der Umfang der Regularien dürfte künftig weiter zunehmen: so veröffentlichte die EU im April 2021 eine Verordnung als ersten Rechtsrahmen für KI [6], die bestehende Regularien ergänzen würde. Der Entwurf teilt KI-Systeme in vier unterschiedliche Risikokategorien [7] ein („unzulässig“, „hoch“, „begrenzt“, „minimal“). Die Risikokategorie definiert dabei Art und Umfang der Anforderungen, die an das jeweilige KI-System gestellt werden. KI-Software, die in die hohe Risikokategorie fällt, muss die strengsten Auflagen erfüllen. <header class=„a-boxheader“ data-collapse-trigger=„“>Checkliste Model Governance</header><div class=„a-boxtarget a-boxcontent a-inline-textboxcontent a-inline-textboxcontent-container“ data-collapse-target=„“> Der Einsatz von Machine Learning bringt Verantwortung und Verpflichtungen mit sich. Um diesen Anforderungen nachzukommen, benötigt ein Unternehmen Prozesse, durch die es <ul class=„boxtitel“>
die Zugriffe auf ML-Modelle kontrolliert
Richtlinien/gesetzliche

Vorgaben umsetzt
die Interaktionen mit den ML-Modellen und deren Ergebnisse verfolgt
festhält, auf welcher Grundlage ein Modell erzeugt wurde Model Governance [8] bezeichnet diese Prozesse in ihrer Gesamtheit <h4>Checkliste:</h4><ul class=„boxtitel“>
Vollständige Modelldokumentation oder Berichte. Dazu gehört auch das Reporting der Metriken durch geeignete Visualisierungstechniken und Dashboards
Versionierung aller Modelle zur Herstellung von Transparenz nach außen (Erklär- und Reproduzierbarkeit)
Vollständige Datendokumentation zur Gewährleistung hoher Datenqualität und Einhaltung des Datenschutzes
Management von ML-Metadaten
Validierung von ML-Modellen (Audits)
Laufendes Überwachen und Protokollieren von Modellmetriken Dazu zählen folgende Aspekte [9]: Robustheit, Sicherheit, Genauigkeit (Accuracy), Dokumentation und Protokollierung sowie angemessene Risikobewertung und Risikominderung. Weitere Anforderungen sind die hohe Qualität der Trainingsdaten, Diskriminierungsfreiheit, Nachvollziehbarkeit, Transparenz, menschliche Überwachung sowie die Erforderlichkeit einer Konformitätsprüfung und der Nachweis der Konformität mit der KI-Verordnung durch eine CE-Kennzeichnung (siehe Kasten „Plan it Legal“). Beispiele für die ML-Systeme dieser Kategorie sind private und öffentliche Dienstleistungen (wie die Bonitätsprüfung) oder Systeme, die in der Schul- oder Berufsausbildung eingesetzt werden, um über den Zugang zu Bildung und den beruflichen Werdegang einer Person zu entscheiden (beispielsweise bei der automatisierten Bewertung von Prüfungen). <header class=„a-boxheader“ data-collapse-trigger=„“>Plan it Legal: KI-Verordnung und Konformität</header><div class=„a-boxtarget a-boxcontent a-inline-textboxcontent“ data-collapse-target=„“><figure class=„a-inline-textboximage-container“><figcaption class=„a-caption a-caption–textbox“>(Bild: Marian Weyo/Shutterstock.com)</figcaption></figure> Die Konformität von HRKI mit der KI-Verordnung wird die Voraussetzung für die Vermarktung in der EU werden. Sie lässt sich über eine CE-Kennzeichnung nachweisen. Die EU wird zudem Standards verabschieden, bei deren Einhaltung die Konformität mit der Verordnung anzunehmen ist. Für die umfassenden Tests, die nach der KI-Verordnung anfallen, sollen die zuständigen Behörden „Sandboxing Schemes“ entwickeln, also Vorgaben für sichere Testumgebungen. Die Konformitätsprüfung für KI beruht auf einer ex-ante-Sicht, hat aber gleichwohl Ähnlichkeiten mit der Datenschutzfolgenabschätzung nach der DSGVO. Mehr Informationen hierzu finden sich im Blogeintrag von Dr. Benhard Freund bei planit.legal: „Das KI-Gesetz der EU – Entwurf und Diskussionsstand [10]“. <h3 class=„subheading“ id=„nav_konformität_für1“>Konformität für europäische KI-Auflagen erreichen</h3> Da die Verordnung nicht nur für in der EU ansässige Unternehmen und Einzelpersonen gelten soll, sondern für jedes Unternehmen, das KI-Dienste innerhalb der EU anbietet, hätte das Gesetz einen ähnlichen Anwendungsbereich wie die DSGVO. Die Verordnung muss sowohl vom EU-Parlament gebilligt werden als auch die Gesetzgebungsverfahren

der einzelnen Mitgliedsstaaten passieren. Wenn das EU-Parlament die Verordnung billigt und sie die legislativen Prozesse der EU-Staaten passiert, tritt das Gesetz frühestens 2024 in Kraft. Dann müssen Hochrisikosysteme während der Entwicklung eine Konformitätsbewertung für KI-Auflagen durchlaufen, um das KI-System in einer EU-Datenbank registrieren zu lassen. Im letzten Schritt ist eine Konformitätserklärung notwendig, sodass KI-Systeme die notwendige CE-Kennzeichnung erhalten, damit ihre Anbieter sie in den Verkehr bringen können. Wichtig ist außerdem, dass Regulierung nicht der einzig ausschlaggebende Aspekt für Model-Governance-Prozesse ist. Denn auch Modelle, die in schwächer regulierten Kontexten im Einsatz sind, kommen an Model Governance nicht vorbei [11]. Neben der Erfüllung gesetzlicher Vorgaben müssen Unternehmen wirtschaftliche Einbußen und Reputationsverluste ebenso abwenden wie juristische Schwierigkeiten. ML-Modelle, die einer Marketing-Abteilung Informationen über die Zielgruppe liefern, können im Betrieb an Präzision verlieren und eine falsche Informationsgrundlage für wichtige Folgeentscheidungen bereitstellen. Somit stellen sie ein finanzielles Risiko dar. Model Governance wird also nicht nur zur Erfüllung rechtlicher Vorgaben, sondern auch zur Qualitätssicherung von ML-Systemen und zur Minderung unternehmerischer Risiken benötigt. <h3 class=„subheading“ id=„nav_model2“>Model Governance als Herausforderung</h3> Die sich abzeichnenden EU-Vorgaben, bestehende Regelungen und Unternehmensrisiken machen es notwendig, Model-Governance-Prozesse von Beginn an zu implementieren. Die Bedeutung von Model Governance ergibt sich für viele Unternehmen allerdings oft erst dann, wenn ML-Modelle in die Produktion gehen und in Einklang mit gesetzlichen Regelungen stehen sollen. Dazu kommt, dass der abstrakte Charakter rechtlicher Vorgaben Unternehmen vor die Herausforderung der praktischen Umsetzung stellt: So geben nach der bereits zitierten Algorithmia-Studie [12] 56 Prozent der Befragten die Implementierung von Model Governance als eine der größten Herausforderungen an, um ML-Anwendungen langfristig erfolgreich in Produktion zu bringen. Dazu passen auch die Zahlen der „State of AI in 2021“-Studie mit Blick auf die Risiken Künstlicher Intelligenz: 50 Prozent der befragten Unternehmen geben die Einhaltung gesetzlicher Vorschriften als Risikofaktor an, andere hoben Mängel bei Erklärbarkeit (44 Prozent der Befragten), Reputation (37 Prozent), Gerechtigkeit und Fairness (30 Prozent) als relevante Risikofaktoren hervor. <h3 class=„subheading“ id=„nav_audits_als3“>Audits als standardisierte Prüfprozesse im Model-Governance-Framework</h3> Ein wichtiger Bestandteil von Model Governance sind Audits [13] als Werkzeuge, um zu prüfen, ob KI-Systeme den Unternehmensrichtlinien, Branchenstandards oder Vorschriften entsprechen. Dabei gibt es interne und externe Audits. Die im Artikel „Ethik und Künstliche Intelligenz: ein neuer Umgang mit KI-Systemen [14]“ auf Heise von der Verfasserin besprochene Studie Gender Shades ist ein Beispiel für einen externen Auditprozess: Sie prüfte Gesichtserkennungssysteme großer Anbieter hinsichtlich ihrer Genauigkeit bezüglich des Geschlechtes und der Ethnie und konnte dabei eine abweichende Präzision des Modells je nach Ethnie und Geschlecht feststellen. Dieser Blick von außen ist aber limitiert, da externe Prüfprozesse nur Zugang zu Modellergebnissen, aber nicht zu den zugrundeliegenden Trainingsdaten oder Modellversionen besitzen. Das sind wertvolle Quellen, die Unternehmen in einem internen Auditprozess einbeziehen müssen. Diese Prozesse sollen eine kritische Reflexion

über die potenziellen Auswirkungen eines Systems ermöglichen. Zunächst sind jedoch an dieser Stelle Grundlagen über KI-Systeme zu klären. <h3 class=„subheading“ id=„nav_besonderheiten4“>Besonderheiten von KI-Systemen</h3> Um KI-Software prüfen zu können, ist es wichtig zu verstehen, wie Machine Learning funktioniert [15]: Maschinelles Lernen besteht aus einer Reihe von Methoden, die Computer verwenden, um Vorhersagen oder Verhaltensweisen auf der Grundlage von Daten zu treffen und zu verbessern. Um diese Vorhersagemodelle aufzubauen, müssen ML-Modelle eine Funktion finden, die zu einer bestimmten Eingabe eine Ausgabe (Label) erzeugt. Dafür benötigt das Modell Trainingsdaten, die zu den Eingabedaten die jeweils passende Ausgabe enthalten. Dieses Lernen trägt die Bezeichnung „überwachtes Lernen“. Im Trainingsprozess sucht das Modell mithilfe mathematischer Optimierungsverfahren eine Funktion, die den unbekannten Zusammenhang zwischen Ein- und Ausgabe so gut wie möglich abbildet. Ein Beispiel für eine Klassifizierung wäre eine Sentimentanalyse, die untersuchen soll, ob Tweets positive oder negative Stimmungen (Sentiments) enthalten. In diesem Fall wäre ein Input ein einzelner Tweet, und das dazugehörige Label das codierte Sentiment, das für diesen Tweet festgelegt wurde (−1 für ein negatives, 1 für ein positives Sentiment). Im Trainingsprozess lernt der Algorithmus mit diesen annotierten Trainingsdaten, wie Eingabedaten mit dem Label zusammenhängen. Nach dem Training kann der Algorithmus dann neue Tweets selbstständig einer Klasse zuordnen. <h3 class=„subheading“ id=„nav_komplexere5“>Komplexere Komponenten im Machine-Learning-Bereich</h3> Somit lernt ein ML-Modell die Entscheidungslogik im Trainingsprozess, statt die Logik mit einer Abfolge von typischen Wenn-Dann-Regeln explizit im Code zu definieren, wie es in der Softwareentwicklung typisch wäre. Dieser grundlegende Unterschied zwischen traditioneller und KI-Software führt dazu, dass sich Methoden des klassischen Softwaretestens nicht direkt auf KI-Systeme übertragen lassen. Das Testen verkompliziert sich dadurch, dass zusätzlich zum Code die Daten und das Modell selbst hinzukommen, wobei alle drei Komponenten sich gemäß dem Change-Anything/Change-Everything-Prinzip gegenseitig bedingen (hierzu mehr unter „Hidden Technical Debt in Machine Learning Systems [16]“). Unterscheiden sich beispielsweise die Daten im produktiven System von den Daten, mit denen ein Modell trainiert wurde (Distribution Shifts), kommt es zum Leistungsabfall des Modells (Model Decay). In diesem Fall muss ein Modell schnell mit frischen Trainingsdaten trainiert und re-deployed werden. Erschwerend kommt hinzu, dass das Testen von KI-Software ein noch offenes Forschungsfeld [17] ohne Konsens und ohne Best Practices ist. <h3 class=„subheading“ id=„nav_ethische6“>Ethische Prinzipien als nicht-funktionale Eigenschaften</h3> Die relevanten Testaspekte von KI-Software lassen sich in funktionale und nicht-funktionale Eigenschaften einteilen. Correctness als funktionale Eigenschaft lässt sich durch Metriken wie Accuracy und Precision/Recall mathematisch direkt erfassen [18]. Sie geben an, wie hoch die Übereinstimmung zwischen den Vorhersagen des trainierten Modells und den tatsächlichen Predictions ist (Gold Standard). Dazu gibt es etablierte Validierungsverfahren wie die Kreuzvalidierung, die durch Isolation der Testdaten über eine Datenstichprobe prüft, wie gut das trainierte Modell die richtigen Modellergebnisse (Labels) für neue Daten vorhersagt. Nicht-funktionale Eigenschaften entsprechen den ethischen Prinzipien [19] wie Fairness, Datenschutz, Interpretierbarkeit, Robustheit und

Sicherheit. Anders als funktionale Eigenschaften können sie nicht auf einen breiten Fundus standardisierter Metriken und Praktiken aus dem Bereich des maschinellen Lernens zurückblicken. Auch hier besteht die Herausforderung darin, dass das Testen nicht-funktionaler Eigenschaften von KI-Software (noch) nicht standardisiert ist. Erschwerend kommen Abwägungen zwischen verschiedenen Eigenschaften hinzu: Fairness verringert die Accuracy [20] und umgekehrt. Metaphorisch lässt sich KI-Software als Kraftwerk bezeichnen: Ein funktional einwandfreier, reibungsloser Betrieb heißt nicht, dass das Kraftwerk der Umwelt keinen Schaden zufügt. Der störungsfreie Ablauf entspricht den funktionalen, der Schutz der Umwelt den nicht-funktionalen Kriterien. Die Metapher zeigt, dass es für funktionale und nicht-funktionale Eigenschaften verschiedene Prüfprozesse braucht. Für Erstere sind Best Practices aus dem ML-Fundus anwendbar, für Letztere hingegen braucht es noch Forschungsarbeit. Im weiteren Verlauf soll es hier um den Aspekt der Fairness, eine Beschreibung von Konzepten sowie Teststrategien und um die Hervorhebung von Model Governance gehen, die die Transformation ethischer Prinzipien wie Fairness in der praktischen KI-Softwareentwicklung unterstützt. <h3 class=„subheading“ id=„nav_wie_entsteht7“>Wie entsteht „Unfairness“?</h3> Zunächst ist die Frage interessant, wie Ungerechtigkeit (Unfairness) überhaupt entsteht [21]. Die Regel ist dabei einfach: Was die Modelle lernen, manifestiert sich in den Trainingsdaten. Im überwachten Lernen bestehen Trainingsdatensätze aus den Eingabedaten und dazugehörigen Labels. Wenn die Datenlabels Bias enthalten, wird das Modell diese Grundeinstellung übernehmen und von Anfang an lernen. Daher ist es wichtig, die Labels ausreichend zu überprüfen. Bias kann sich aber auch inhärent aus den Daten, nicht nur aus den Labels ergeben: Enthalten die Trainingsdaten an sich bereits Bias, greift der Algorithmus ihn ebenfalls auf. Dieses Problem besteht etwa bei umfangreichen, mit großen Datenmengen aus dem Internet trainierten Sprachmodellen. Es ließ sich nachweisen, dass die Leistungsstärke eines Modells mit der Stärke eines stereotypen Bias korreliert: Mit steigender Präzision nimmt auch der Bias zu. Auch ein geringer Stichprobenumfang bei Minderheitsgruppen kann zu einer Homogenisierung des Lernprozesses des Modells zugunsten der Mehrheitsgruppen führen, beispielsweise durch mehr Fotos männlicher als weiblicher Gesichter in den Trainingsdaten. Neben den Daten spielen auch im Trainingsprozess verwendete Merkmale (Features) eine Rolle. Kann das Modell nicht ausreichend viele Merkmale nutzen, erschwert das dem Algorithmus, den Zusammenhang zwischen Ein- und Ausgabe zu lernen. Aus diesem Grund reagierte IBM mit Diversity in Faces als Versuch, die Diversität der Fotos in den Trainingsdaten zu erhöhen [22]. Und schließlich können Merkmale „Stellvertreter“ für ausgeschlossene sensible Attribute sein: Auch wenn geschützte Attribute bei der Entscheidungsfindung nicht explizit verwendet werden, können sie implizit beteiligt sein, wenn sie mit den ausgeschlossenen Merkmalen korrelieren. <header class=„a-boxheader“ data-collapse-trigger=„“>Fairness in KI-Systemen</header><div class=„a-boxtarget a-boxcontent a-inline-textboxcontent a-inline-textboxcontent-container“ data-collapse-target=„“> Das Ziel der Gewährleistung von Fairness ist der Schutz sensibler Attribute wie Geschlecht, Religionszugehörigkeit oder sexueller Orientierung vor unfairer algorithmischer Entscheidungsfindung. Das Recht auf Diskriminierungsfreiheit ist im EU-Rechtsentwurf für KI-Systeme der hohen Risikokategorie explizit verbrieft. Während sich die Ungerechtigkeit bei der Gender-Shades-Studie leicht intuitiv erfassen lässt, besteht nun die Herausforderung darin, den abstrakten Begriff der Fairness objektiv, metrikbasiert und möglichst skalierbar zu definieren. <h3

class=„subheading“ id=„nav_definitionen8“>Definitionen für Fairness und Ableitungen von Teststrategien</h3> Welche Audits und welche Metriken bieten sich an, um Fairness zu testen? Die bereits bekannte Konsenslücke klafft auch für die Definition von Fairness [23] auseinander. Erschwerend kommt hinzu, dass die Vielfältigkeit der verschiedenen Ursachen für Fairness zeigt, dass sich Fairness nicht mit einer simplen Metrik oder Teststrategie herstellen lässt – Fairness-Audits müssen Teil der Model-Governance-Prozesse sein, die die Qualitätssicherung der Trainingsdaten und des Modells sicherstellen. Dazu kommt, dass die verschiedenen Anwendungsfälle für KI zu vielfältig sind, als dass es eine gut generalisierende One-Size-Fits-All-Lösung geben könnte. Die Frage, wie Fairness gemessen und nachgewiesen werden kann, lässt sich also nicht nur an einer simplen Metrik festmachen. Dennoch soll es zunächst um konkrete Möglichkeiten gehen, Fairness quantitativ zu erfassen, bevor diese Audits in das Model-Governance-Framework eingebettet werden. Statistische Ansätze bieten die am leichtesten messbaren Definitionen von Fairness, und sie bilden gleichzeitig die Grundlage für weiterführende Ansätze. Zur Quantifizierung von Fairness lassen sich statistische Metriken nutzen. Von diesen Messgrößen leiten sich Definitionen ab, die sich auf die Ausgabe von Modellen konzentrieren. Fairness lässt sich aufgrund ähnlicher Fehlerquoten der Ausgaben für unterschiedlich sensible demografische Gruppen [24] definieren. Entsprechend ist ein Algorithmus dann fair, wenn Gruppen, die auf der Grundlage sensibler Attribute ausgewählt werden, die gleiche Wahrscheinlichkeit von vorteilhaften Entscheidungsergebnissen haben („Group Fairness [25]“). <h3 class=„subheading“ id=„nav_gleichheit_der9“>Gleichheit der Gesamtgenauigkeit</h3> Zudem lässt sich untersuchen, ob die Genauigkeit des Modells für verschiedene Subgruppen gleich ist (Gleichheit der Gesamtgenauigkeit). Am Beispiel einer Kreditwürdigkeitsprüfung wäre diese Definition von Fairness dann erfüllt, wenn die Wahrscheinlichkeit für Personen männlichen und weiblichen Geschlechts gleich ist, dass Antragsstellende mit einem tatsächlich guten Kreditscore als kreditwürdig eingestuft werden und dass solchen mit einem schlechten Kreditscore die Kreditwürdigkeit abgesprochen wird, ohne Ansehen der Geschlechtszugehörigkeit. Für das Testen statistischer Ansätze sind bereits erste Lösungen verfügbar: Fairness Indicators von TensorFlow [26] ist eine Bibliothek, die die Berechnung häufig identifizierter Fairness-Metriken mit verbesserter Skalierbarkeit auf großen Datensätzen und Modellen bietet. Darüber hinaus unterstützt Fairness Indicators die Auswertung der Verteilung von Datensätzen und der Modellleistung über verschiedene Benutzergruppen sowie die Berechnung statistisch signifikanter Unterschiede auf der Basis von Konfidenzintervallen. <h3 class=„subheading“ id=„nav_statistische10“>Statistische Ansätze und Counterfactual Fairness</h3> Zwar sind statistische Ansätze gut messbar, sie können jedoch zu kurz greifen. Fairness lässt sich nicht allein durch ähnliche Fehlklassifizierungsquoten erklären, insbesondere, wenn alle anderen Attribute mit Ausnahme des sensiblen Attributs ignoriert werden. Beispielsweise könnte ein KI-System zur Kreditwürdigkeitsprüfung demselben Anteil männlicher und weiblicher Bewerber eine positive Bewertung zuweisen – statistische Ansätze würden das Modell dann als gerecht beurteilen. Doch wenn die männlichen Bewerber zufällig ausgewählt wurden, während weibliche Bewerber schlicht diejenigen sind, die die meisten Ersparnisse haben, wäre Fairness nicht

gegeben. Similarity-based Measures stellen nicht die Modellergebnisse und Fehlklassifizierungsquoten, sondern den Prozess der Entscheidungsfindung sowie die Verwendung von Merkmalen im Trainingsprozess in den Vordergrund. Daraus lässt sich „Fairness through Unawareness“ als Konzept für Fairness ableiten: Algorithmen können als fair gelten, wenn geschützte Attribute aus den Trainingsdaten ausgeschlossen worden sind. In unserem Beispiel bedeutet dies, dass geschlechtsspezifische Merkmale nicht für das Training des Modells verwendet werden, sodass Entscheidungen nicht auf diesen Merkmalen beruhen können. Doch auch dieser Ansatz hat Einschränkungen: Das Ausschließen geschützter Attribute reicht nicht aus, da andere, ungeschützte Attribute Informationen enthalten können, die mit den ausgeschlossenen geschützten Attributen korrelieren („Counterfactual Fairness [27]“). In diesem Fall wäre das ursprünglich ausgeschlossene Attribut implizit in anderen Attributen enthalten und würde den Entscheidungsprozess indirekt beeinflussen (siehe auch Kasten „Kontrafaktisch testen“). <header class=„a-boxheader“ data-collapse-trigger=„“>Kontrafaktisch testen</header><div class=„a-boxtarget a-boxcontent a-inline-textboxcontent a-inline-textboxcontent-container“ data-collapse-target=„“> Causal-Reasoning-Ansätze stützen sich auf Werkzeuge der Kausalinferenz. Die Definition der kontrafaktischen Fairness basiert auf der Intuition, dass eine Entscheidung gegenüber einer Person dann fair ist, wenn sie in der tatsächlichen Welt und in einer kontrafaktischen Welt, in der die Person einer anderen demografischen Gruppe angehört, gleich ist. Damit ist Counterfactual Fairness dann gegeben, wenn sich eine Prediction nicht ändert, obwohl das geschützte Attribut in das kontrafaktische Gegenteil verkehrt wird. Beispielsweise müsste die Entscheidung für oder gegen die Kreditwürdigkeit einer Person gleich ausfallen, wenn das Attribut von „männlich“ in „weiblich“ verändert wird. <h3 class=„subheading“ id=„nav_unfairness_mit11“>Unfairness mit manipulierten Daten aufdecken</h3> Adversiales Testen ist eine gängige Strategie, die zur Aufdeckung von Schwachstellen einen böswilligen Angriff auf ein System simuliert. Beim adversialen Testen erhält das Modell Eingabedaten, die mit kleinen, absichtlichen Merkmalsmanipulationen versehen sind. Auf diese Weise wird getestet, ob das Modell für speziell zugeschnittene Eingabedaten unerwünschte Vorhersagen trifft [28]. Die Manipulation der Eingabedaten ist domänenspezifisch und kann von Analysen algorithmischer Unfairness inspiriert sein. Die Idee, die Reaktion eines Modells auf Eingabedaten zu testen und auf diese Art von Bias zu quantifizieren, findet sich mittlerweile auch in Frameworks. So lässt sich mit dem Benchmarking-Datensatz StereoSet prüfen [29], ob und wie stark die Vorurteile von Sprachmodellen bezüglich Geschlecht, Ethnie, Religion und Profession ausgeprägt sind: Entwicklerinnen und Entwickler können fertig trainierte Sprachmodelle einreichen, um diskriminierende Entscheidungsfindung in Sprachmodellen zu messen und gleichzeitig die Sprachmodellierungsleistung zu berücksichtigen. StereoSet betrachtet die Gesamtleistung des Modells als gut, wenn das Modell in der Lage ist, den Zielkonflikt zwischen Genauigkeit und Fairness abzuschwächen und so ein genaues Verständnis natürlicher Sprache bei gleichzeitiger Minimierung von Verzerrungen zu gewährleisten. Frameworks wie diese können zwar eine gute Leitlinie sein, ersetzen aber nicht das individuelle Testen, das fest in einem Prüfungsprozess eingebettet sein muss. <h3 class=„subheading“ id=„nav_warum_audits12“>Warum Audits alleine nicht ausreichen</h3> Die verschiedenen Prüfprozesse sind wichtig – sie allein genügen aber nicht, um Fairness zu gewährleisten. Vielmehr kommen Audits als eine der ersten Möglichkeiten in Betracht,

um Probleme zu identifizieren. Sie müssen Teil des Model-Governance-Frameworks sein und sollten es ergänzen. Für sich allein besitzen sie hingegen keine Aussagekraft – nur ein ganzheitlicher Ansatz kann alle Aspekte berücksichtigen, die für Fairness eine Rolle spielen. Neben der Validierung funktionaler und nicht-funktionaler Anforderungen, die die hier beschriebenen Auditprozesse zum Testen von Fairness enthalten kann, ist sauberes Dokumentieren eine weitere wichtige Komponente im Model-Governance-Framework. Dokumentationen sollten bereits in der ersten Phase des ML-Lebenszyklus, der Entwicklung, begonnen werden (mehr dazu im „Practictioners' Guide to MLOps [30]“). In der Development-Phase geht es um den Aufbau einer robusten und reproduzierbaren Trainingsprozedur, die aus Datenverarbeitungs- und Modellaufbauschritten besteht. Dieser Aufbauprozess ist experimentell und iterativ, wobei wichtige Informationen über Daten (Auswahl und Definition von Features, Aufteilung der Trainings-, Validierungs- und Testdaten, Schema und Statistiken), Modelle (verschiedene getestete Modelle) und Parameter festzuhalten sind (Experimental Tracking). <h3 class=„subheading“ id=„nav_nach_dem13“>Nach dem Training: KI-Modelle evaluieren</h3> Nach dem Aufbau der Trainingsprozedur gilt es, entwickelte Modelle hinsichtlich funktionaler und nicht-funktionaler Eigenschaften zu evaluieren (an dieser Stelle sind die beschriebenen Auditstrategien zum Testen von Fairness relevant). Die Ergebnisse der Evaluation und alle Informationen über den Aufbauprozess der Trainingsprozedur fließen in die Dokumentation ein, die zusätzlich eine Erklärung des Use-Case-Kontextes, eine High-Level-Erklärung des Algorithmus, Modellparameter, Anweisungen zur Reproduktion des Modells und Beispiele für das Training der Algorithmen sowie Beispiele für das Treffen von Predictions durch den Algorithmus enthalten sollte. Die Dokumentation lässt sich durch Toolkits wie Model Cards und Data Sheets praktisch unterstützen. Data Sheets halten fest, welche Mechanismen oder Verfahren für die Datenerhebung verwendet wurden oder ob ethische Überprüfungsverfahren (Audits) stattgefunden haben. Model Cards [31] ergänzen Data Sheets [32] und informieren über die Art der Modellerstellung, die bei der Entwicklung getroffenen Annahmen sowie über Erwartungen bezüglich des Modellverhaltens bei verschiedenen kulturellen, demografischen oder phänotypischen Gruppen. <h3 class=„subheading“ id=„nav_unternehmenspoli14“>Unternehmenspolitik als Schlüssel für ethische KI</h3> Vollständige Dokumentation schafft Reproduzierbarkeit und Transparenz nach außen. Nach dem Deployment muss diese Sichtbarkeit (Observability) im produktiven System gegeben sein [33]. Hier spielt zum einen die Versionierung von Modellen und Datensätzen eine wichtige Rolle. Die Versionierung dient der Wahrung des Unveränderlichkeitsgrundsatzes der Modelle, sodass alle Modelle sich ohne Datenverluste und Veränderung reproduzieren lassen. Damit ist auch gewährleistet, dass eine Model Prediction der Modellversion, die sie produziert hat, zugeordnet werden kann. Zum anderen muss ein Monitoring-System die Leistung des produktiven Modells kontinuierlich überwachen und relevante Metriken in einem Report zusammenfassen sowie visualisieren. Diese Werte aus dem Model-Logging sollten in Metriken aufbereitet und in Dashboards zu Protokollierungs-, Analyse- und Kommunikationszwecken visualisierbar sein. Wird im Monitoring der Leistungsabfall eines Modells (Model Decay) festgestellt, muss das Modell mit neuen Trainingsdaten trainiert und dann re-deployed werden. <h3 class=„subheading“ id=„nav_empfehlung15“>Empfehlung: Audits vor jedem neuen Deployment</h3> Vor jedem neuen Deployment sollten erneut Audits stattfinden, um ethische, rechtliche oder geschäftliche Risiken zu kontrollieren. Ethische Prinzipien wie Fairness sind auf

jeder Ebene der Softwareentwicklung zu berücksichtigen, unter anderem bereits bei der Datenbeschaffung. Fairness lässt sich nicht mit einer simplen Metrik oder Teststrategie herstellen: Es braucht eine entsprechend ausgerichtete Unternehmenspolitik, die das anerkennt. Ohne Model Governance sind KI-Systeme hinsichtlich der Einhaltung gesetzlicher Vorgaben sowie der Minderung des unternehmerischen Risikos unkalkulierbar. <header class=„a-boxheader“ data-collapse-trigger=„“>Young Professionals schreiben für Young Professionals</header><div class=„a-boxtarget a-boxcontent a-inline-textboxcontent“ data-collapse-target=„“><figure class=„a-inline-textboximage-container“></figure> Isabel Bär studiert Data Engineering am Hasso-Plattner-Institut [34] und arbeitet als Werkstudentin bei INNOQ [35]. Sie beschäftigt sich mit Fragen rund um den langfristig erfolgreichen Einsatz von Künstlicher Intelligenz (KI), wozu insbesondere MLOps und das Implementieren von Model Governance gehören. () <hr /> URL dieses Artikels:
<small><code>https://www.heise.de/-6352125</code></small> Links in diesem Artikel:
<small><code>[1] https://www.heise.de/hintergrund/Alpine-js-Das-Schweizer-Taschenmesser-fuer-dynamische-Weboberflaechen-6177628.html</code></small>
<small><code>[2] https://www.heise.de/hintergrund/Developer-Experience-Gluckliche-Entwickler-schreiben-besseren-Code-6150890.html</code></small>
<small><code>[3] https://www.heise.de/hintergrund/Ethik-und-Kuenstliche-Intelligenz-ein-neuer-Umgang-mit-KI-Systemen-6056059.html</code></small>
<small><code>[4] https://www.heise.de/developer/young-professionals-6065678.html</code></small>
<small><code>[5] https://algorithmia.com/blog/new-report-discover-the-top-10-trends-in-enterprise-machine-learning-for-2021</code></small>
<small><code>[6] https://digital-strategy.ec.europa.eu/en/policies/european-approach-artificial-intelligence</code></small>
<small><code>[7] https://ec.europa.eu/info/strategy/priorities-2019-2024/europe-fit-digital-age/excellence-trust-artificial-intelligence_de</code></small>
<small><code>[8] https://algorithmia.com/blog/model-governance</code></small>
<small><code>[9] https://germany.representation.ec.europa.eu/news/fur-vertrauenswurdige-kunstliche-intelligenz-eu-kommission-legt-weltweit-ersten-rechtsrahmen-vor-2021-04-21_de</code></small>
<small><code>[10] https://planit.legal/das-ki-gesetz-der-eu-entwurf-und-diskussionsstand/</code></small>
<small><code>[11] https://algorithmia.com/blog/model-governance</code></small>
<small><code>[12] https://algorithmia.com/blog/new-report-discover-the-top-10-trends-in-enterprise-machine-learning-for-2021</code></small>
<small><code>[13] https://dl.acm.org/doi/pdf/10.1145/3351095.3372873</code></small>
<small><code>[14] https://www.heise.de/hintergrund/Ethik-und-Kuenstliche-Intelligenz-ein-neuer-Umgang-mit-KI-Systemen-6056059.html</code></small>
<small><code>[15] https://christophm.github.io/interpretable-ml-book/what-is-machine-learning.html</code></small>
<small><code>[16] https://papers.nips.cc/paper/2015/file/86df7dcfd896fcaf2674f757a2463eba-Paper.pdf</code></small>
<small><code>[17] https://ieeexplore.ieee.org/abstract/document/8920538</code></small>
<small><code>[18] https://www.aaai.org/Papers/Workshops/2006/WS-06-06/WS06-06-003.pdf</code></small>
<small><code>[19] https://ieeexplore.ieee.org/document/9000651</code></small>
<small><code>[20] https://people.mpi-sws.org/~gummadi/papers/process_fairness.pdf</code></small>
<small><code>[21] https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2477899</code></small>
<small><code>[22] https://www.ibm.com/blogs/research/2019/01/diversity-in-faces/</code></small>
<small><code>[23] https://arxiv.org/pdf/1906.10742.pdf</code></small>
<small><code>[24] https://people.mpi-sws.org/~gummadi/papers/process_fairness.pdf</code></small>
<small><code>[25] https://people.ece.ubc.ca/mjulia/publications/Fairness_Definitions_Explained_2018.pdf</code></small>
<small><code>[26] https://github.com/tensorflow/fairness-indicators</code></small>
<small><code>[27] https://papers.nips.cc/paper/2017/file/a486cd07e4ac3d270571622f4f316ec5-Paper.pdf</code></small>
<small><code>[28] https://christophm.github.io/interpretable-ml-book/adversarial.html</code></small>
<small><code>[29] https://stereoset.mit.edu/</code></small>
<small><code>[30] https://services.google.com/fh/files/misc/practitioners_guide_to_mlops_whitepaper.pdf</code></small>
<small><code>[31] https://arxiv.org/pdf/1810.03993.pdf</code></small>
<small><code>[32] https://arxiv.org/pdf/1803.09010.pdf</code></small>
<small><code>[33] https://www.oreilly.com/library/view/the-framework-for/9781098100483/ch01.html</code></small>
<small><code>[34] https://hpi.de/</code></small>
<small><code>[35] https://www.innoq.com/de/</code></small>
<small><code>[36] mailto:sih@ix.de</code></small>
 Copyright © 2022 Heise Medien </html>