Ethische Prinzipien und die Regulierung künstlicher Intelligenz

© Sandra Hermannsen

Politiker*innen und Ethikforscher*innen stellen die Frage, wie künstliche Intelligenz (KI) gesellschaftlich und ethisch verantwortungsvoll entwickelt, eingesetzt und genutzt werden kann. Dr. Simon Hirsbrunner arbeitet am Internationalen Zentrum für Ethik in den Wissenschaften (IZEW) der Universität Tübingen und beschäftigt sich in seiner Forschung mit verschiedenen Thematiken angewandter KI-Ethik. *mediendiskurs* sprach mit ihm.

JOACHIM VON GOTTBERG IM GESPRÄCH MIT SIMON HIRSBRUNNER

„Kritische Stimmen aus der KI-Entwicklergemeinde sind natürlich grundsätzlich zu begrüßen."

Mehr als 350 Wissenschaftler, darunter KI-Entwickler wie Sam Altman von OpenAI, aber auch der frühere KI-Entwickler bei Google Geoffrey Hinton haben in einem Brandbrief vor den Gefahren der KI gewarnt: „Es sollte global priorisiert werden, das Risiko der Auslöschung durch KI zu verringern – auf einer Stufe mit anderen Risiken für die gesamte Gesellschaft, wie etwa Pandemien und Nuklearkrieg." Auch Elon Musk hat im März ein halbjähriges Moratorium in der Entwicklung von KI gefordert.

Kritische Stimmen aus der KI-Entwicklergemeinde sind natürlich grundsätzlich zu begrüßen. Dennoch ist die Art und Weise, wer hier wie wovor warnt, nicht so wirklich sinnvoll und manchmal auch etwas unehrlich. Die Unterzeichnenden beschwören die Möglichkeit eines Endes der Menschheit durch KI herauf. Damit fällt der Brandbrief auf philosophische Grundsatzfragen der KI-Ethik zurück, von denen wir uns eigentlich schon vor einem Jahrzehnt verabschiedet haben – Fragen zum Szenario der Singularität oder Superintelligenz, also einer zukünftigen Situation, in der künstliche Intelligenz dem Menschen ebenbürtig oder sogar überlegen ist. In der aktuellen KI-Ethik und öffentlichen Debatte dazu hat man sich aber in den letzten Jahren eher auf die konkreten und teils schwerwiegenden Probleme konzentriert, die mit bereits operationalisierten KI-Systemen einhergehen – also mit der Thematik der algorithmischen Diskriminierung, dem Einsatz fehleranfälliger Anwendungen, den schweren Eingriffen in die Privatheit von Bürger*innen und mit der fehlenden Transparenz aktueller KI-Systeme.
Der Brandbrief wirkt deshalb wie ein Störfeuer, welches den Fokus der öffentlichen Debatte weg von Themen führt, die für die KI-Unternehmen gefährlich werden können – namentlich dem Ruf nach konsequenter und robuster KI-Regulierung. Zusätzlich lässt der Brief die Führungsriege sehr gut dastehen – als (einzige) Retter der Menschheit, die uns vor einer Roboter-Apokalypse bewahren können.
Lassen Sie uns also lieber über aktuelle ethische Herausforderungen sprechen, die sich gegenwärtig durch die Verankerung von KI in vielen Lebensbereichen ergeben.

Gerne. Die Grundlage für eine funktionsfähige KI sind massenhafte Daten. Wo kommen die her?

Es gibt viele unterschiedliche KI-Verfahren, und deshalb gibt es auch sehr unterschiedliche Datenquellen. Wir haben gerade auf der re:publica in Berlin eine Veranstaltung zu KI im Datenjournalismus durchgeführt. Da ging es um die Frage, wie Daten für journalistische Zwecke verfüg- und nutzbar gemacht werden können und welche Funktionen hier KI übernehmen mag. Im Kontext des investigativen Journalismus ging es hier um die Analyse heterogener Massendaten, die den Journalist*innen beispielsweise im Zuge der Ermittlungen zu den Pandora- und Panama-Papers zugespielt wurden – also um Finanzdaten in Tabellenform, Chatprotokolle von Beteiligten, Verträge in Form von PDF-Dateien. In anderen Anwendungsfeldern von KI sehen die Daten aber komplett anders aus.
Es kann sich dabei auch um Datenformate handeln, die gar nicht für eine Interpretierbarkeit durch den Menschen strukturiert, sondern ausschließlich für Maschinen lesbar sind. Das ist beispielsweise der Fall bei vielen KI-basierten Prozessoptimierungsalgorithmen im Kontext der sogenannten Industrie 4.0.

*Kann ich als Nutzer*in KI-basierter Anwendungen der Weiternutzung meiner Daten widersprechen? Sam Altman, CEO von OpenAI, hat das der italienischen Regulierungsbehörde versprochen, nachdem ChatGPT aus Datenschutzgründen in Italien verboten worden war.*

Da gibt es rechtlich gesehen Unterschiede zwischen Daten mit Personenbezug und Daten, die anonymisiert wurden. Die Nutzung personenbezogener Daten wird in der EU durch die Datenschutz-Grundverordnung (DSGVO) geregelt. Hier haben wir als Nutzer*innen ein Anrecht auf informiertes Einverständnis („informed consent"). Allerdings wird diese Nutzungserlaubnis meist im Rahmen des Registrierungsprozesses für digitale Dienstleistungen abgefragt und entsprechend auch oft erteilt. Eine zweite Frage bezieht sich auf Nutzerdaten, die anonymisiert wurden und entsprechend nicht mehr unter die DSGVO fallen. Auch die Nutzung dieser anonymisierten Daten kann problematisiert werden. Zum einen muss infrage gestellt werden, ob Daten überhaupt effektiv vor einer Deanonymisierung geschützt werden können. Dies

„Bisher spielt der Datenschutz bei ChatGPT und Co. keine besonders wichtige Rolle."

wird in der Wissenschaft von verschiedenen Seiten angezweifelt. Zum anderen werden die anonymisierten Daten zur Profilierung sozialer Gruppen verwendet, beispielsweise zum Zweck individualisierter Werbung. Auch wenn diese Profilierung auf anonymisierten Daten aufbaut, hat sie letztendlich doch eine starke Auswirkung auf Autonomie von Individuen, da diese nicht entscheiden können, zu welchen algorithmischen Gruppen sie gehören.

Die Frage ist: Nehmen die Firmen meine Einstellungen wirklich ernst?

Bisher spielt der Datenschutz bei ChatGPT und Co. jedenfalls keine besonders wichtige Rolle. Die Firmen, deren Software wir vor allem verwenden, kommen auch aus einem Kulturkreis, in dem der Wert des Datenschutzes nicht auf dem gleichen Niveau angesiedelt ist wie in Deutschland. Derzeit fressen Anwendungen wie ChatGPT und Google Bard mit großem Appetit alle Daten, die ihnen von einem globalen Pool von Nutzer*innen großzügig zur Verfügung gestellt werden.

KI-Systeme zur Personalauswahl vermuten auf der Grundlage historischer Daten, Chefpositionen seien immer mit alten weißen Männern besetzt worden, und wählen diese entsprechend für zukünftige Stellenbesetzungen aus. Kann man solche Mechanismen algorithmischer Diskriminierung beheben?

Einerseits muss man sagen, dass die Informatik viele Ideen entwickelt hat, um diskriminierenden Verzerrungen („biases") in KI-Systemen entgegenzuwirken. Manchmal lassen sich diese Probleme algorithmischer Diskriminierung also wirklich sehr effektiv durch statistische Verfahren beheben – beispielsweise, indem KI-Modelle mit diverseren Daten gefüttert werden, die auch Informationen zu unterrepräsentierten sozialen Gruppen (Frauen, Schwarze Menschen, Menschen mit Migrationshintergrund) miteinschließen.
Andererseits sollte Fairness auch nicht nur auf statistische Repräsentativität enggeführt werden. Viele mit KI in Bezug stehende Formen der Diskriminierung finden außerhalb der Modelle und Daten statt – beispielsweise, wenn Auswahlverfahren bei schlecht bezahlten Jobs an die KI outgesourct werden, während Kandidat*innen für Kaderpositionen immer noch die „Ehre" haben, mit menschlichen Personaler*innen sprechen zu dürfen. Insofern ist Fairness im Kontext von KI immer als ethischer Wert zu verstehen, dem man sich annähern will, und nicht die Eigenschaft eines technischen Systems. Es ist problematisch, wenn Firmen behaupten, ihr KI-basiertes Produkt sei „komplett objektiv und fair", nur weil mittels technischer Verfahren diskriminierende Verzerrungen statistisch aufgelöst wurden.

In dem Film Minority Report *oder dem Roman von Sebastian Fitzek* Das Joshua-Profil *kann ein Programm verbrecherische Profile prognostizieren: Die Personen können verhaftet werden, bevor sie eine Tat begehen.*

Dieses Predictive Policing ist ein sehr kontroverses Anwendungsfeld für KI. Vor fünf oder zehn Jahren gab es verschiedene Softwareangebote in den USA, die angeblich solche Vorhersagen machen konnten. Aber es stellte sich heraus, dass diese technischen Systeme ethnische Diskriminierungen richtiggehend beflügeln. Ein Beispiel: Das System bewertete die Gefährlichkeit verschiedener Gegenden in Städten anhand der dort durchgeführten Festnahmen. Wo in der Vergangenheit viele Festnahmen registriert worden waren, dorthin schickte das System präventiv Polizeikräfte, um so zukünftige Verbrechen zu verhindern. Was jedoch nicht mitgedacht wurde: Wenn die Polizei in den USA in Gegenden mit einer mehrheitlich Schwarzen Bevölkerung mit großem Polizeiaufgebot aufläuft, führt dies automatisch zu mehr Festnahmen, weil die Präsenz von Polizei dort nicht als Garant für Sicherheit, sondern als Mittel der Unterdrückung wahrgenommen wird. Je mehr Polizei dort ist, desto mehr Festnahmen entstehen, worauf das technische System immer weitere Polizeibeamte losschickt und es damit wieder zu mehr Festnahmen kommt. So entsteht ein Teufelskreis algorithmischer Diskriminierung.
Daneben etabliert derartige Software aber auch einfach ein System durchgehender Überwachung, welches die Autonomie des Individuums maßgeblich infrage stellt.
Diese möglichen Auswirkungen wurden ja in Filmen wie *Minority Report* eindrücklich thematisiert.

„KI verschärft eine Reihe von Problematiken, mit denen sich die Menschheit schon seit einer Ewigkeit herumschlägt."

Hatespeech, Fake News oder Social Bots, die Wahlen manipulieren sollen: Kann man mit KI desinformierende Inhalte und Akteure aufspüren und von Social-Media-Plattformen verbannen?

Bei den Social-Media-Plattformen wird die Erkennung problematischer Inhalte schon jetzt algorithmisch gelöst, etwa durch die Uploadfilter von YouTube. Wichtig ist dabei der Human-in-the-Loop: Algorithmen können Bearbeitende zwar auf problematische Inhalte hinweisen, Menschen sollten aber dennoch das letzte Wort haben, ob nun Inhalte gelöscht oder Nutzer*innen gesperrt werden. Im Moment funktioniert das aus Kostengründen viel zu automatisch und entsprechend fehlerhaft.

*Stellen Sie sich folgendes Szenario vor: Studierende erstellen ihre Masterarbeit mit ChatGPT und der*die Hochschullehrer*in weiß nicht, was und wie sie das nun bewerten soll…*

Die Bildungscommunity hat da meines Erachtens ein wenig überreagiert. Vor vielen Jahren hatten wir eine ähnliche Befürchtung: Kinder und Jugendliche würden alles Wissen nur noch von der Wikipedia kopieren – und verdummen. Damals waren die Informationen auf Wikipedia teilweise auch wirklich sehr rudimentär und entsprechend nicht vertrauenswürdig. Seither hat sich jedoch ein effektives Kuratierungs- und Prüfsystem etabliert, welches die Qualität, Glaubhaftigkeit und Ausgewogenheit der Informationen in Wikipedia-Artikeln sicherstellt. Heute sind viele, wenn auch nicht alle Artikel ausgesprochen hochwertig und natürlich jeweils aktueller als der *Brockhaus*. Anders gesagt: Die Menschen haben sich die neue Wissenstechnologie erfolgreich zu eigen gemacht und sie ihren Bedürfnissen angepasst. Ähnliches müssen wir mit Blick auf KI auch in Angriff nehmen.

Thematiken, die aus ethischer Sicht im Blick behalten werden sollten, sind die potenziellen Abhängigkeiten und Ungleichheiten, die sich durch die Nutzung von KI im Bildungsbereich ergeben. Akteuren im Bildungsbereich muss bewusst gemacht werden, dass Systeme wie ChatGPT und Midjourney mehr sind als das Eingabeinterface, welches auf Knopfdruck überzeugende Medieninhalte ausspuckt. Diese Systeme greifen im Hintergrund auf eine riesige globale Infrastruktur zurück, die nicht nur aus digitalen Daten besteht, sondern auch aus Menschen und materiellen Ressourcen; beispielsweise Armeen von Clickworkern im Globalen Süden, die zu teils miserablen Arbeitsbedingungen Trainingsdaten annotieren. Ebenso Nutzer*innen, welche die Systeme unbewusst mittels Interaktion weiter trainieren. Hinzu kommt der Einsatz seltener Ressourcen, die in den Systemen verbaut werden (seltene Erden) und der hohe Stromverbrauch. Durch die Nutzung von KI-Systemen werden auch Bereiche wie der Schulbetrieb oder die Universitätslehre von dieser riesigen Infrastruktur abhängig. Die Abhängigkeit schafft potenziell neue Ungleichheiten, etwa wenn bestimmte Gebiete (ländliche Regionen, der Globale Süden) keinen Zugang zu Hochgeschwindigkeits-Internetanschlüssen und damit auch nicht zu KI-Anwendungen haben.

Lassen Sie es mich so zusammenfassen: KI verschärft eine Reihe von Problematiken, mit denen sich die Menschheit schon seit einer Ewigkeit herumschlägt – beispielsweise Diskriminierung marginalisierter Gruppen, Ungleichheit zwischen Weltregionen, Eingriffe in die Privatsphäre durch die Mächtigen, undurchsichtige staatliche und privatwirtschaftliche Entscheidungsprozesse. Da diese Problematiken nicht grundsätzlich neu sind, verfügen wir aber auch über ein effektives Instrumentarium, um sie adressieren zu können – durch die Formulierung ethischer Standards und Best Practices, einen umfassenden Aufbau von Kompetenzen in unterschiedlichen Lebens- und Arbeitsbereichen sowie durch die Verabschiedung und Durchsetzung eines konsequenten Rechtsrahmens. Der Europäische KI-Regulierungsentwurf („AI Act"), der noch dieses Jahr verabschiedet werden soll, ist hier sicherlich ein Schritt in die richtige Richtung. Da KI-Technologie sich aber so rasant weiterentwickelt, sind ein robuster Rechtsrahmen und die Formulierung ethischer Richtlinien keine Selbstläufer. Ethische KI ist eine gesamtgesellschaftliche Herausforderung, die wir nur gemeinsam und mit stetigem Einsatz meistern können.