DENK FABRIK DIGITALE ARBEITSGESELLSCHAFT

Präsentation: KI als Entscheider*in
Prof. Dr. Katharina Zweig

Das gesamte Redemanuskript von Prof. Dr. Katharina Zweig

Ja, ich freue mich sehr, hier zu sein. Für uns ist das ein spannender Moment. Ich bin jetzt bestimmt seit drei Jahren in der Forschung an der Frage dran: Was ist sie denn nun diese KI? Warum reden wir so viel darüber? Und ich glaube, dass Bundestagspräsident Dr. Schäuble das in seiner Rede zur konstituierenden Sitzung der Enquete-Kommission Künstliche Intelligenz, in der ich auch Mitglied bin, ganz gut zusammengefasst hat. Er sagte, die Künstliche Intelligenz gilt vielen als neue Zauberformel des technischen Fortschritts. Und dann holt er ein bisschen aus, was sie alles können wird und sagte unter anderem, sie wird dichten. Und sie wird belohnen und strafen. Ich glaube das, sind die zwei Ängste, mit denen wir es hier zu tun haben. Sie wird dichten und sie wird richten. Wofür steht das Dichten? Das steht da voll für die Sorge, dass uns die KI in unseren menschlichen Eigenschaften ersetzen könnten oder auch Tätigkeiten ersetzen könnte. Und das Richten, gerichtet werden von einer Maschine. Das ist etwas, vor dem viele Leute Respekt haben und deswegen möchte ich mich auf diesen Punkt konzentrieren, die Frage danach, wie gehen wir jetzt um damit, wenn es KI-Systeme gibt, die Entscheidungen übernehmen? Was ich damit meine, ist das folgende. Es gibt einen Algorithmus, der kann gelernte Regeln beinhalten. Und der bekommt Menschen als Input und scored diese Person oder klassifiziert sie. Ist in diesem Raum noch jemand, der das gruselig findet? Das sah vor zwei Jahren noch ganz anders aus. Da hätten sehr viele gestreckt und das ist ja schon mal ein gutes Zeichen. Denn eigentlich kennen wir das natürlich, das kennen wir von der Schufa, das kennen wir von der Kfz-Versicherung. Aber in diese Systeme haben wir auch einiges an Vertrauen. Wir wissen, was der gesellschaftliche Zweck dieser Systeme ist, wir wissen einigermaßen, welche Daten hineingehen und das Ergebnis hängt von unserem eigenen individuellen Verhalten ab. Und wenn Menschen über uns entscheiden, dann können wir uns auch auf gewisse Fakten verlassen. Menschliche Entscheidungen. Die gibt es natürlich in allen Formen und Farben. Das kann gehen von dem netten Obstverkäufer, der mir vielleicht einen Apfel schenkt, an den wir aber von seiner Ausbildung her nicht viele Ansprüche stellen. Das kann gehen über die Bürokraft, die vielleicht eine Entscheidung über mich trifft. Dann ist hier schon der Elektriker und der macht etwas brandgefährliches. Natürlich muss der gut ausgebildet sein, damit klar ist, dass mein Haus nicht in Flammen aufgeht. Und dann haben wir Architekten, die stabile Häuser bauen sollen und nicht zuletzt haben wir unsere Richterinnen und Richter. Und sie sehen schon, ich habe hier so eine kleine Dramaturgie drin. Es geht um immer wichtigere Entscheidungen.

Und deswegen stellen wir auch immer höhere Ansprüche an unsere menschlichen Entscheider und Entscheiderinnen. Natürlich wollen wir diese Ansprüche übertragen auf die KI. Jetzt muss man aber auf der anderen Seite sagen, so richtig ins Hirn gucken können wir unseren menschlichen Entscheidern auch nicht. Das heißt ein bisschen, eine Blackbox sind Sie für uns und für andere natürlich auch. Aber wir können diese Anforderungen stellen an die Ausbildung von menschlichen Entscheiderinnen. Die darin erworbenen Qualifikationen vielleicht auch an spezifische Inhalte, Fortbildungen oder an ein gewisses Erfahrungs-Level. Wir können Entscheidungen zusätzlich absichern, wenn sie uns besonders wichtig sind durch Widerspruchsmöglichkeiten, durch unabhängige Gutachten, durch das Hinzuziehen mehrerer Experten und Expertinnen und Hierarchien von Entscheidern und Entscheiderinnen, bei denen wir uns orientieren können. Das heißt, mit der Wichtigkeit der Entscheidung, und das ist eigentlich der Anwendungskontext, es geht hier nicht in erster Linie um das KI-System, sondern wo es nachher genau eingesetzt wird, steigen in aller Regel im Allgemeinen die Anforderungen an die Ausbildung und die Möglichkeiten, an die Ausbildung des Entscheiders oder der Entscheiderin und die Möglichkeiten des Widerspruchs. Was passiert jetzt, wenn KI entscheidet? Ist denn da wirklich etwas fundamental anders? Und ja, das ist es, denn so wie KI-Systeme gebaut werden, sind ganz schön viele Hände daran beteiligt. Und das Demonstrieren wir immer als sogenannte lange Kette der Verantwortlichkeiten und die Entwicklung eines KI-Systems beginnt dort, wo viel Expertise ist im Algorithmen-Design. Das macht dann so jemand wie ich. Dann gibt es vielleicht eine zweite Gruppe von Personen, die diese Algorithmen, die mal theoretischer Natur sind, implementiert, also als Software Package anbietet. Und tatsächlich sind wir genau für diese beiden Schritte sehr gut ausgebildet, wir beweisen mathematisch, dass ein Algorithmus das tut, was er soll. Und unsere Studierenden haben gelernt, ein Algorithmus zu nehmen und den in die Praxis zu übertragen. Gleichzeitig gibt es Personen oder Institutionen, die Daten sammeln und dann gibt es da diesen neuen Beruf und der ist noch nicht so gut reguliert, das sind im Moment Physiker, Mathematiker, Informatiker, manchmal auch Domänen-Spezialisten, die sich Statistik drauf geschafft haben und die wählen nun eine Methode aus des maschinellen Lernens und wählen die Trainingsdaten aus. Dann gibt es jemanden, der ein sogenanntes Qualitäts- und Fairnessmaß ausbildet. Denn so, wie das funktioniert, ist das ein sehr interaktiver Prozess. Man wählt eine Methode aus, man wählt die Daten aus und dann guckt man mal, wie gut die Entscheidungen der KI sind auf sogenannten Testdaten. Und um zu gucken wie gut die sind, braucht man dieses Qualitäts-Maß und zusätzlich vielleicht ein Fairnessmaß, um Diskriminierungen zu vermeiden. Solange diese Werte nicht gut genug sind, verändert man etwas an der Methode und da sind vielmehr Knöpfchen und Regler als sie es sich normalerweise vorstellen. Es ist also ein sehr iterativer Prozess. Irgendwann sind wir damit zufrieden und dann gibt es ein statistisches Modell mit Entscheidungsregeln. Dann gibt es wieder eine andere Person, die jetzt Daten einfüttert in diesen Algorithmus, irgendjemand der das Ergebnis interpretiert und dann daraus eine Handlung ableitet und vielleicht passiert das auch vollautomatisch.

Vielleicht haben wir dann auch noch ein Feedback, sodass das System ständig weiter lernen kann. Na ja und Sie sehen schon, hier sind sehr viele Personen an der Ausbildung des KI-Systems beteiligt, an der Auswahl der Erfahrung, die dem System zuteilwerden. Und deswegen ist es eben auch schwieriger als bei Menschen festzustellen, ob sie fachlich geeignet sind, um eine Entscheidung zu machen. Also was ändert sich durch den Einsatz von KI? Im Gegensatz zum menschlichen Entscheider und Entscheiderinnen? Es ist eine unklare Ausbildung, die per se erst mal nicht transparent ist. Dadurch wird es unklar, welche Qualifikation das System eigentlich hat, um eine Entscheidung zu treffen. Es gibt fehlendes Kontextwissen, beim Menschen wissen wir immer, dass die grundlegend verstehen, wie die Welt funktioniert, insbesondere, wenn sie bestimmte Ausbildungslevel erreicht haben. Können wir davon ausgehen, dass dieses Hintergrundwissen da ist? Eine KI nach heutigem Zuschnitt wurde immer nur auf eine Sache trainiert. Und wenn bestimmte Kontexte nicht unter ihren Erfahrungen waren, dann wird die Maschine sie nicht haben. Natürlich gibt es auch ein großes Plus. Deswegen wollen wir sie ja vielleicht verwenden, wenn sie gut genug sind. Sie können sehr viel mehr Erfahrung machen als wir es in einem Menschenleben jemals tun könnten. Es gibt mit ihnen die Möglichkeit nach viel mehr Erfahrungsmustern zu suchen, als wir es als Menschen jemals tun könnten und selbst sehr kleine Muster, die wir als Menschen gar nicht betrachten würden, sind von den statistischen Verfahren des maschinellen Lernens durchaus noch verwertbar. Das kann gut sein, das kann schlecht sein, denn es kann sich auch einfach um ein statistisches Rauschen handeln. Und ganz zuletzt, warum machen wir das alles? Naja, mit so einer Maschine haben wir die Möglichkeit, das skalieren zu lassen und Entscheidungen plötzlich für nahezu alle Personen zu treffen und auch das kann gut sein, kann effizienter machen, kann aber auch schwierig sein, weil wir dann plötzlich nur noch einen Entscheider haben und eben keine Hierarchien mehr. Wie bewerten wir das jetzt? Wie sollten wir eine solche KI bewerten? Wir glauben, dass es notwendig ist, den Anwendungskontext mitzunehmen. Denn es gibt sehr viele KI-Systeme, die so allgemein sind, dass man sie in sehr verschiedenen Kontexten einsetzen kann. Dazu gehören sehr einfache KI-Systeme, zum Beispiel Produktempfehlungs-Systeme. Es macht einen Unterschied, ob Sie mir damit meine neue Sommerkleidung anbieten oder ob sie herauskriegen wollen, welche politische Werbung sie an wen verteilen wollen. Es kann auch bei einem System, das nicht in so groß unterschiedlichen Anwendungskontexten verwendet wird, aber für zwei Zielgruppen einen großen Unterschied machen. Denken Sie an YouTube, das ihnen das nächste Video vorschlägt. Da gibt es andere Dinge zu betrachten für 18-Jährige als für 13-Jährige als für Zwei- bis Dreijährige, die gar nicht wegkönnen. Daher glauben wir, dass man das Schadenspotenzial betrachten muss, das sich zusammensetzt aus einem Schadenspotenzial für Individuen und einem möglicherweise darüber hinausgehenden Schadenspotenzial für die Gesellschaft als Ganzes. Auf einer zweiten Achse muss man sich ansehen, wie groß der Grad der Abhängigkeit der bewerteten von einer solchen Entscheidung ist.

Das heißt, man kann dabei nachsehen, wie viele Anbieter gibt es, welche Widerspruchs- und Wechselmöglichkeiten gibt es. Und wenn die klein sind und das Schadenspotenzial hoch ist dann müssen wir besser hinein gucken können in die Qualifikation, in die Erfahrung, in die Ausbildung. Wie kann man das machen? Wie kann man denn feststellen, dass man überhaupt von einem Schaden betroffen ist? Dazu gibt es in der Literatur verschiedene Transparenz-und Nachvollziehbarkeits-Mechanismen, von denen ich hier einige gelistet habe. Man kann fordern, Transparenz über die Daten, auf denen trainiert wurde und die dann nachher auch eingegeben werden, über deren Qualität und Vollständigkeit, über die verwendete Methode des maschinellen Lernens über das gewählte Qualität- und Fairnessmaß und wie es abgeschnitten hat, in welcher Form von Test. Natürlich wurde auch viel gefordert, dass man eine Einsicht in den Code bräuchte. Dies ist wirklich fast niemals sinnvoll und fast niemals notwendig, aber ich habe es hier der Vollständigkeit halber mit gelistet. Bei den Nachvollziehbarkeits-Mechanismen geht es darum, dass ich das nicht unbedingt glauben muss, was mir gesagt wird, sondern dass ich es selber nachprüfen kann, dass ich nachvollziehen kann, welche Daten verwendet wurden. Ob sie wirklich qualitativ so hochwertig sind und vollständig sind, wie gesagt wurde, dass ich selber Tests durchführen kann und die Tests, die schon gefahren wurden, nachvollziehen kann und dass sie schlussendlich auch die Bewertung von Qualität und Fairness nachvollziehen kann. Das ist eine nicht vollständige Liste, aber sie sehen schon, so viel mehr als das ist es auch gar nicht und die Frage ist jetzt, was davon muss welches KI-System in welcher Anwendung wirklich erfüllen? Und dafür schlagen wir eine einfache Klassifizierung in fünf Klassen vor, von denen wir glauben, dass es eine große grüne gibt, wo wir keinerlei Einsicht brauchen in die Ausbildung, in die Qualifikation, weil es sich um fast kein Schadenspotenzial handelt, zum Beispiel, wenn sie ihre Nasenrücken vermessen lassen, damit sie ihre Brille angepasst bekommen, solange die Daten da bleiben, sind das zwar persönliche Daten, Schadenspotenzial nicht vorhanden. Eine Postdoc Analyse kann zeigen, wenn es doch einen Schaden gab, was genau schiefgelaufen ist und danach würde man das System sicherlich neu bewerten. Dann haben wir drei Klassen, in denen wir verschiedene Kombinationen dieser Transparenz- und Nachvollziehbarkeits-Mechanismen sehen und eine sehr kleine, wo wir tatsächlich sagen würden, hier sollten Maschinen ganz grundsätzlich nicht über Menschen entscheiden. Wie könnte das genau aussehen? Wir haben hier eine vorläufige Zuteilung der verschiedenen Transparenz- und Nachvollziehbarkeits-Mechanismen in die verschiedenen Klassen. Und sie haben es heute Morgen schon gehört. Ja, horizontal, das will irgendwie eigentlich keiner. Das ist aber ein horizontaler Ansatz. Ich werde viel gefragt von Industrievertretern, muss das sein, können wir nicht in den Sektoren bleiben? Aber tatsächlich ist es eben so, wenn KI-Systeme für so viele verschiedene Anwendungskontexte verwendet werden, kann man das im Vorhinein nicht sagen, ob es ein kritischer Bereich ist oder nicht. Deswegen schlagen wir vor, eine sehr kleine sehr dünne, sehr schnelle horizontale Regulierung einzuführen, in der man sich einteilt in eine dieser fünf Klassen.

Dann werde ich immer wieder gefragt, wie viele wird denn das betreffen? Ich kann Ihnen keine genauen Zahlen geben, aber wir haben in den letzten Tagen 130 KI-Systeme per Hand klassifiziert und 60 fallen ganz klar davon, 60 Prozent fallen ganz klar davon in die Klasse Null und von denen, die nicht in Klasse Null fallen, sind 50 Prozent Medizinprodukte. Weitere 20 Prozent im Bereich kritischer Infrastrukturen oder der Mensch-Roboter-Interaktion. Dass wir da ein bisschen Einsicht brauchen in die Qualität der Entscheidungen, ist glaube ich offensichtlich und auch für die betreffenden Branchen nicht wirklich überraschend. Also, warum ist diese KI-Regulierung komplex? Weil wir es hier mit vielen verschiedenen Anwendungskontexten haben und die entsprechenden Berufe über die letzten 1500 Jahre, angefangen bei den Ständen und Ausbildungsregularien reguliert haben. Und jetzt müssen wir all das noch einmal anfassen und noch einmal überprüfen, ob es im Bereich der KI-Systeme immer noch Sinn macht oder nicht, denn KI-Systeme ersetzen oder unterstützen Personen in diesen sehr unterschiedlich stark regulierten Prozessen und Positionen. Aber das hat auch sein Gutes. Denn vielleicht ist es doch gar nicht so komplex, denn im Endeffekt ist es gar nicht so viel Neues. Wir haben die sozialen Prozesse mit hohem Schadenspotenzial schon heute reguliert. Medizinprodukte sind seit Ewigkeiten unter entsprechenden Regularien, weil viele Aushandlungsprozesse zur Verortung von Verantwortlichkeit ebenfalls schon existieren und nur noch transferiert werden müssen und weil auch die meisten Gruppen deren Rechte vielleicht beeinträchtigt sein könnten oder die geschädigt werden könnten, auch schon Schutzinstitutionen haben und diese müssen wir jetzt natürlich ein bisschen aufrüsten. Also ich gehe mit dieser positiven Note raus. Ich glaube, wir müssen jetzt aus der Praxis jetzt den Transfer leisten. Deswegen habe ich mich auch gefreut, dass das erste Panel hieß, From Principles to Praxis. Wir haben ganz viele Prinzipien, wir haben auch die eigentlichen Handlungsanleitungen schon da und jetzt ist die Frage, wie wir das in die Praxis bekommen. Und deswegen Liebes KI-Observatorium, wünsche ich Dir zu Deinem Geburtstag, dass du zu einem Fixpunkt werden mögest in dieser Dynamik. Vielen Dank!