Wissenswert: Crashkurs KI (6): Explainability - Von der Black Box zum Glaskasten.

Link zur hr-info-Seite mit dem Audio

- Crashkurs KI (3): Müssen wir Arbeit völlig neu denken?
- Crashkurs KI (4): Kampf der KI-Großmächte
- Crashkurs KI (5): Fairness
- Crashkurs KI (7): Lügende KI

Müssen wir wissen, was Maschinen denken?

hr-iNFO "Wissenswert",1.8.2020

Vor über hundert Jahren staunten die Menschen über den "Klugen Hans", Ein Pferd, das rechnen kann, und lesen, sogar ein bisschen schreiben. War es Betrug - oder war das wirklich eine neue Art der Intelligenz? Tatsächlich war der "Kluge Hans" zweifellos intelligent - aber es stellte sich heraus, dass diese Intelligenz ganz anders war als vermutet. Und dieser Fall vom Beginn des 20. Jahrhunderts hat Bedeutung für unseren heutigen Blick auf das intelligente Verhalten von Computersystemen, denn auch dort ist nicht klar, was eigentlich passiert. Was denkt sich die KI eigentlich? Antworten auf diese Frage hören Sie in der nächsten halben Stunde.

Anmerkungen und Quellen als Fußnoten im Text.

Jan Eggers, Hessischer Rundfunk, Juli 2020

Musik: untergeekDE - https://soundcloud.com/untergeekde/sets/machine-learning-music-for-a-radio-feature-on-ai

Kluger Hans (1) Wiehern...Morgen

Wilhelm von Osten ist stolz - Er hat Erstaunliches erschaffen: Eine nicht menschliche Intelligenz, die zählen kann, addieren, subtrahieren, Wurzeln ziehen. Gegenstände erkennen und benennen, und sogar buchstabieren - also auch lesen und schreiben. Wilhelm von Osten hat diese Intelligenz trainiert. Davon schließlich versteht er was, der Herr Kaiserliche Volksschullehrer der Mathematik im Ruhestand. Er hat den Klugen Hans geschult[1]: einen achtjährigen Orlow-Traberhengst; ein Pferd, das offensichtlich nicht nur rechnen kann, sondern sogar schreiben - mit einer Art Schreibmaschine für Hufe.

Kluger Hans (1) Jawohl, richtig

Kann das sein? Selbst eine 13-köpfige kaiserliche Forscherkommission findet keinerlei Anzeichen für Betrug. Der junge Forscher Oskar Pfungst darf die Kommission im Jahre 1904 begleiten. Er beschreibt die Szene so[2].

(p) Schweinhardt Zitat für Jan Eggers

Der Schauplatz war ein gepflasterter Hof im Norden Berlins, inmitten hoher Mietskasernen. Auf dem Hof ein alter Mann von 60 bis 70 Jahren, das weiße Haar von einem Schlapphut bedeckt. zu seiner Linken ein stattlicher Rapphengst… die in deutscher Sprache ohne besondere Betonung an ihn gerichteten Fragen beantwortete er fast ausnahmslos richtig. Hatte er eine Frage verstanden, so gab er dies sogleich durch Nicken zu erkennen, das Gegenteil durch Kopfschütteln.

---

[1] Kurt Tutschek hat für den Wiener "Standard" die Geschichte vom "Klugen Hans" viel schöner erzählt als ich - und außerdem das Buch von Oskar Pfungst verlinkt, aus dem ich unten zitiere.

Wer etwas Niederländisch versteht, kann sich die Geschichte auch anschauen: Die niederländische Schulfunkreihe "Gedrag en Wetenschap" hat die Geschichte vom Klugen Hans in Szene gesetzt - den Film finden Sie bei Youtube: https://www.youtube.com/watch?v=3J5e0hT-UkA

[2] Oskar Pfungst, Das Pferd des Herrn von Osten. Ein Beitrag zur experimentellen Tier- und Menschenpsychologie (S. 18f.) - https://archive.org/details/daspferddesherr00stumgoog

Möglicherweise schütteln jetzt auch Sie den Kopf beim Zuhören. Es sollte doch um KI gehen, nicht um Zirkuskunststückchen! Aber die Geschichte vom "Klugen Hans" hat eine Pointe, die die Informatiker heute noch umtreibt.

Vom "Klugen Hans" zur KI, die irrte

Das Pferd war ohne Zweifel recht intelligent und gelehrig - für ein Pferd. Aber es konnte weder rechnen, lesen noch schreiben. Das hat der wiederum (für einen Menschen) recht intelligente Oskar Pfungst durch Versuche herausgefunden: Das Pferd achtete auf unwillkürliche kleine Signale seines Trainers. Wilhelm von Osten gab durch seine Körperhaltung zu erkennen, wenn der Pferdehuf in der Nähe der richtigen Lösung war; das Pferd hatte gelernt, diese sekundären Signale zu verstehen und sich gewissermaßen von seinem Besitzer fernsteuern zu lassen - ohne dass der es merkte. Eine erstaunliche Leistung - nur eben kein Beweis für Intelligenz.

Das kann auch mit KI passieren: Ein Team deutscher Forscher hat den Klugen Hans gewissermaßen in Software nachgebaut. Sie konnten nachweisen, dass ein Algorithmus zur Bilderkennung zwar korrekt Pferde von Autos unterscheiden konnte - aber nur, weil alle Pferde-Fotos zufällig unten links einen Copyright-Hinweis hatten - ein sekundäres Signal, das die Maschine sich merkte.[3] Ein anderes Forscherteam hat eine lernende Maschine darauf trainiert, Wölfe von Hunden zu unterscheiden - aber sie haben die KI bewusst in eine Falle gelockt: Die Fotos der Wölfe zeigten allesamt auch Schnee, sodass der Computer letztlich etwas Falsches lernte: Vierbeiner auf Schnee: gleich Wolf, Vierbeiner ohne Schnee: gleich Hund.[4]

Diese Versuche dienen nicht dazu, zu zeigen, wie wenig man Künstlicher Intelligenz, also lernenden Maschinen, trauen kann. Die Forscherinnen und Forscher versuchen eher, ein Problem zu lösen: Wir können zeigen, dass Künstliche Intelligenz zu außergewöhnlichen

---

[3] Lapuschkin, Wäldchen u.a.: "Unmasking Clever Hans predictors and assessing what machines really learn". Nature Communications 10, 2019 https://www.nature.com/articles/s41467-019-08987-4#Abs1

[4] Das Paper dazu: https://www.kdd.org/kdd2016/papers/files/rfp0573-ribeiroA.pdf Vom Projekt, das dahinter steht, wird unten noch mehr die Rede sein.

Erkennungsleistungen fähig ist - aber wir können nicht so richtig erklären, wie sie das tut. Eine KI ist heute in der Regel eine Black Box - eine Maschine, die zwar funktioniert, aber über deren Inneres wir nichts Sinnvolles sagen können.

Explainability - Erklärbarkeit - ist die Forderung, diese Black Box ausleuchten zu können. Sie ist eins der spannendsten Forschungsgebiete im Maschinellen Lernen - und ist eine Antwort auf eine zutiefst menschliche Frage: Sind wir bereit, der KI zu vertrauen? Vertrauen setzt voraus, dass wir verstehen, was die Algorithmen tun - denn erst mit der Erklärung kann man nachsteuern, wenn die KI sich vergaloppiert hat.

Kluger Hans (1) Wiehern k

Warum die Black Box nicht mit uns spricht

An dieser Stelle müssen wir einen Blick darauf werfen, weshalb es so schwierig ist, das Innere einer künstlichen Intelligenz zu erkunden. Nehmen wir das Beispiel von vorhin: Wir wollen einer Maschine beibringen, Hunde und Wölfe zu unterscheiden.

Dazu brauchen wir erst einmal möglichst viele Fotos von Hunden und Wölfen, die wir, die Menschen, sortiert haben - ein Stapel mit Wolfs-Fotos, einen mit Hunde-Fotos. Mit diesen Fotos, die ja letztlich nur aus Pixeln bestehen, wird die Maschine nun trainiert. Die Maschine lernt, ein bestimmtes Muster von Eingangssignalen mit dem Etikett "Hund" oder "Wolf" zu verbinden. Und das funktioniert so ganz anders als bei uns Menschen, viel automatischer.

Wenn wir Wölfe von Hunden unterscheiden, fließt unser Wissen über die Welt mit ein. Die Mathematikerin Hannah Fry schildert in ihrem (lesenswerten) KI-Buch "Hello World"[5], wie ein Mathematikprofessor mit seinem vierjährigen Enkel an einem Wolfshund vorbeikommt. Woher weißt du eigentlich, dass das kein Wolf ist, fragt der Professor. Und der Kleine antwortet:

[5] Hannah Fry, ***Hello World***. Was Algorithmen können und wie sie unser Leben verändern. München, C.H.Beck 2019. Die geschilderte Anekdote steht auf Seite 107.

OT Leine: "Weil er an der Leine ist."

Menschen lernen über Verstehen: Sie entwickeln eine Vorstellung davon, was einen Hund ausmacht und was einen Wolf. Die KI lernt nur einen Zusammenhang zwischen bestimmten Mustern von Eingangssignalen und dem gewünschten Ausgangssignal. Signale, die dann in der Software immer besser gruppiert und miteinander verrechnet werden, bis eine bestimmte Form von Eingangssignalen das Ausgangssignal "Wolf" oder "Hund" produziert.

Der KI-Forscher und -Pionier Judea Pearl hat das so ausgedrückt: Maschinelles Lernen sei vor allem dadurch erfolgreich, indem es uns zeigt, dass bestimmte Aufgaben, die wir für anspruchsvoll gehalten haben, es eigentlich nicht sind.[6]

Pearl ordnet aktuelle KI-Algorithmen und maschinelles Lernen ganz unten in einer Hierarchie ein, die er die "Leiter der Kausalität" nennt: Auf der untersten Ebene beobachten lernende System einfache statistische Zusammenhänge; sie lernen zu assoziieren: Aha, wer die Bohnen dieser Pflanze zu sich nimmt, wird wach. Auf der nächsten Ebene lernen sie, diese Zusammenhänge bewusst zu erproben: Wenn ich wach werden will, kann ich mir dann aus den Bohnen der Pflanze morgens einen Sud brühen - und es muss die richtige Pflanze sein, nicht nur irgendeine Bohne. Die Stufe der Intervention nennt Judea Pearl das. Noch weiter oben auf der Leiter der Kausalität ist die Stufe der Imagination: Was wäre, wenn es einen Super-Kaffee gäbe, der nicht nur wach macht, sondern klug? Wenn ich auf Kaffee grundsätzlich verzichten würde? Wenn Schweine fliegen könnten? Das können schon Vierjährige, und sind damit jeder KI haushoch überlegen.

Jemand, der helfen möchte, KI-Systeme auf die nächste Stufe zu heben, ist Niki Kilbertus. Der junge Informatiker forscht gemeinsam mit seinen Kolleginnen und Kollegen am Max-Planck-Institut für

---

[6] Nachzulesen in der Einleitung zu Pearls Standardwerk: Judea Pearl, ***The Book of Why. The New Science of Cause and Effect***. Penguin Random House UK, 2018 Richtet sich an Nicht-Informatiker und versucht, die Möglichkeiten und Beschränkungen der KI zu umreißen.

Intelligente Systeme in Tübingen[7] - und versucht, Maschinen die Möglichkeit zu geben, auch Wenn-Dann-Zusammenhänge zu erfassen.

> ***NK1 Judea Pearl***
>
> Judea Pearl forscht hauptsächlich an Kausalität und sagt eben, diese Idee, ein Modell im Kopf zu haben, mit dem man Welten simulieren kann, die so nicht passiert sind - die klassische Was-wäre-wenn-Frage, oder wenn ich mich jetzt so entscheiden würde, was würde dann passieren? Was würde mit anderen Entscheidung passieren? Dazu muss man wirklich die kausalen Zusammenhänge verstehen.
>
> Das kann man nicht lernen, wenn man nur auf historische Daten blickt…

...aber nichts anderes KI-Systeme derzeit. In unserem Beispiel: die vom Menschen vorsortierten Wolfs- und Hunde-Foto-Datensätze. Deshalb kann man sie auch nicht fragen, warum sie etwas entscheiden.

Ein Blick in die Black Box

Zusammengefasst: KI, künstliche Intelligenz, ist nicht intelligent. Wenn wir die lernenden Maschinen mit unserem Verstand vergleichen wollen, dann nicht mit menschlichem Verstehen, sondern eher mit erlernten Reflexen. So wie ein Hund auf seinen Namen zu hören lernt - oder ein Pferd wie der "Kluge Hans" den Zusammenhang zwischen unwillkürlichen Bewegungen eines Menschen und einem gewünschten Verhalten erlernen kann.

Anders als bei Pferden, Hunden und Menschen steckt bei den Maschinen einen Menge Mathematik dahinter. Nun sollte man meinen, dass es das leichter macht, den erlernten Zusammenhängen nachzuspüren - denn anders als ein Gehirn können wir einen Rechner anhalten und uns jeden einzelnen Zahlenwert des Programms ausgeben lassen. Nur leider wird das schnell unübersichtlich.

---

[7] Vorstellung des Teams am Lehrstuhl von Prof. Bernhard Schöllkopf und seiner Arbeit: https://www.mpg.de/13640712/auf-fairness-programmiert

Lernende Maschinen, die Bilder erkennen sollen, arbeiten mit neuronalen Netzen - Programmstrukturen, die sich grob an den Verschaltungen von Nervenzellen in Menschen- und Tiergehirnen orientieren. Statt Nervenzellen sind bei den Maschinen in neuronalen Netzen viele kleine Rechenwerke miteinander verschaltet. Aus einer kleinen Zahl Eingangssignale, sagen wir: den Bildpunkten eines Fotos, sollen sie eine Zahl von Ausgangssignalen produzieren. Dazu wird die Einstellung der vielen kleinen Rechenwerke immer wieder angepasst. Ein Mechanismus namens Backpropagation passt die Einstellungen an, ähnlich wie ein Klavierstimmer immer wieder an jeder einzelnen Saite dreht, um das Instrument in Stimmung zu bekommen. Ein Klavier hat 230 Saiten - die Anzahl der Werte in einem neuronalen Netz zur Bilderkennung geht in die Hunderttausende, wenn nicht sogar in die Millionen.

Kersting1 Neuronale Netze

Sie hatten von solchen tiefen neuronalen Netzen gesprochen, mit diesen Millionen und ganz vielen Parametern in kleinen Berechnungseinheiten, die da irgendetwas machen...

Das ist Kristian Kersting. Professor für KI und Maschinelles Lernen in Darmstadt.[8] Auch er und seine Mitarbeiter suchen nach einer Lösung für das Problem der Erklärbarkeit.

Kersting 2

Also das Einfachste ist zu sagen: ich möchte verstehen, wie ein neuronales Netz im Prinzip funktioniert. Ich kann zwar jetzt nicht diese Millionen von Parametern im Kopf haben. Aber ich kann Ihnen sagen: okay, bei dieser Eingabe wird jetzt das in der nächsten Schicht, also in den nächsten Neuronen, in diesem kleinen Einheiten berechnet. Und dann wird wieder das Berechnungen wieder das berechnet. Ist es das , was wir wollen,

[8] Auf diesen Seiten stellt Kristian Kersting sich und seine Arbeit vor: https://www.ml.informatik.tu-darmstadt.de/ - Träger des "Deutschen KI-Preises" der Springer-Publikation Bilanz ist er übrigens auch.

nein, wir wollen ja irgendwie eine Erklärung. Ein Verständnis vielleicht, mit dem wir was anfangen können.

Und für dieses Verständnis nützen einem die inneren Zustände nicht - aber man kann die Maschine dazu bringen, ein wenig genauer zu zeigen, worauf sie besonders schaut.

Kersting 3 rückwärts rechnen

Man kann versuchen, wenn ich eine Eingabe habe, dann rechnet das durch, und es kommt eine Ausgabe raus. Und jetzt kann ich mir überlegen: wie kann ich entscheiden, welche Eingabe relevant war für die Ausgabe? Man lässt das Netzwerk gefühlt rückwärts laufen und fragt sich: kann ich jetzt zum Beispiel sagen, dieser Bereiche im Bild, der war wichtig, um zu sagen, das da was ich dort sehe, ist ein Mensch, und diese andere Bereich, der war dafür nicht so wichtig.

Kristian Kersting setzt dafür unter anderem einen Algorithmus namens LIME von Forschern an der University of Washington ein. Die haben das Beispiel mit der Maschine entwickelt, der Wölfe und Hunde unterscheiden sollte, aber tatsächlich auf den Schnee im Hintergrund schaute statt auf das Tier. Und sie haben, typisch amerikanisch, sogar ein kleines Werbevideo dafür produziert:[9]

LIME 1 Wolves

let's say we have a machine learning system that can detect wolves. We can look at accuracy and at some of the predictions and they may look good...

Ein System, das Wolfsbilder erkennt, und, und wenn wir uns die Erkennungsrate sieht soweit gut aus…

...jetzt können die Forscher mehr Informationen darüber bekommen, wie das zustande kommt. Sie können mit LIME oder anderen Verfahren der Explainable AI das Netzwerk rückwärts laufen lassen und zeigen,

[9] Das Video - zusammen mit Beispiel-Code, mit dem man das Verfahren ausprobieren kann - haben die Forscher ins Netz gestellt: https://github.com/marcotcr/lime

welche Teile der Fotos besonders wichtig für die Bilderkennung sind. Das System produziert dann Fotos, in denen die entscheidenden Bildpunkte farbig markiert sind - bei den Wolfsfotos ist das dann nicht der Wolf, sondern der Hintergrund mit dem Schnee. So kann der Mensch sehen, dass das System zwar vielleicht die richtigen Entscheidungen trifft, aber aus den falschen Gründen.

LIME 2 Detection

If our wolf detection system is actually just detecting snow in the background, we will not trust it, even if the accuracy is high.

Welche Wirkung hat welche Ursache? Die Antwort auf diese Fragen geht weit über akademisches Interesse hinaus.

Metzinger Chemo

Stellen Sie sich vor, Sie sind Patient im Krankenhaus, und die KI sagt, wir würden in ihrem Fall nicht noch eine Chemo machen, sondern wir würden jetzt einfach in die palliativ Therapie übergehen. Auf Deutsch gesagt es ist Zeit für sie zu sterben. Und dann sagen Sie als Patient halt. Ich will aber, dass mir das erklärt wird.

Dieses Beispiel hat der Professor für theoretische Philosophie und Neuro-Ethiker Thomas Metzinger schon einmal an anderer Stelle im Crashkurs KI beschrieben. Ihm ging es dabei um die Ansprüche des Patienten auf Information, aber das Kluger-Hans-Phänomen bedeutet, dass die Antwort noch aus einem anderen Grund wichtig ist: Was, wenn die KI, die das Röntgenbild analysiert, mit Bildern trainiert wurde, in denen noch Spuren der Anmerkungen des Radiologen sichtbar waren? Können wir sicher sein, dass die KI allein den Tumor analysiert - und nicht irgendwelche anderen Hinweisen gefolgt ist wie z.B. diesen Bleistiftspuren?

Das setzt voraus, dass die Rolle von KI in der Medizin genau beschrieben ist: sie liefert sachliche Analysen, ohne Ansehen der

betroffenen Person. Frage an den KI-Forscher Kristian Kersting: wie müsste sich eine Medizin-KI dem Patienten gegenüber erklären?

Kersting 4 Ärzte

Vorab, meine Frau ist Ärztin. Deswegen darf ich auch etwas über Ärzte sagen, aber es gibt klassische Studien in der Psychologie, das Experten wie zum Beispiel Chefärztinnen und -Ärzte ihr Wissen auch internalisieren. Also: sie können nicht mehr so einfach darauf zugreifen, und dann entstehen so Situation, die wir alle vielleicht kennen. Man fragt, ja, aber warum haben Sie diese Therapie jetzt gemacht?, und man kriegt eher so ne Antwort: Na ja, das hat aber auch bei dem gut geklappt, und wir haben gute Erfahrungen gemacht…

Wir versuchen, eine Abstraktionsebene zu finden, die es Ihnen immer noch verständlich macht. Und aber auch dem Experten irgendwie. Und ich denke, das muss auch das Ziel bei diesen Maschinen sein - und es geht vielmehr darum, ob wir diesen Maschinen vertrauen können.

Da ist es wieder, das Schlüsselwort: Vertrauen. Ein Begriff, der so gar nicht in die mathematische Welt der Informatik zu passen scheint - die aber zum Teil der Forschung geworden ist.

Wann sind wir Menschen bereit, den Entscheidungen einer KI zu vertrauen? Dazu hat Kristian Kersting ein interessantes Experiment gemacht.[10] Er hat Studentinnen und Studenten vor ein KI-Programm gesetzt, das immer wieder beim Nutzer nachfragte: Habe ich diesen Datensatz richtig beurteilt? Ist das Ergebnis richtig, und sind die richtigen Daten berücksichtigt?

Kersting 5a Erklärung

[10] Stefano Teso, Kristian Kersting: "Why Should I Trust Interactive Learners?" Explaining Interactive Queries of Classifiers to Users" https://arxiv.org/abs/1805.08578 (Preprint)

Jetzt kann die Maschine dem Benutzer, also den Menschen sagen: guck mal, das ist meine Erklärung. Also meinetwegen unten das Copyright wird hervorgehoben…

...so wie im Beispiel, bei dem die KI Pferde-Fotos erkennen sollte...

Kersting 5b Feedback

...und dann kann der Mensch sagen: du, richtige Vorhersage, aber echt blöde blöde Erklärung. Und dieses Signal, dieses Feedback kann man mathematisch wieder ausnutzen, dass diese Maschine weiter lernen kann.

Tatsächlich sorgte dieser Austausch nicht nur für eine besser funktionierende KI, sondern vor allem auch für mehr Vertrauen bei den Testpersonen, gerade am Anfang, wenn die KI zu Beginn ihres Trainings eher schlechte Vorhersagen trifft.

Mechanismen, mit denen eine KI beim Menschen nachfragt - dieser Idee traut der KI-Forscher Kristian Kersting einiges zu. Denn schließlich muss, um zum Medizin-Beispiel zurückzukehren, der Arzt der KI vertrauen, dann kann er dem Patienten erklären, dass er das tut, worauf wiederum der Patient dem Arzt vertrauen können sollte. Und dieses Vertrauen entsteht erst durch den Austausch.

Kersting 6 Forschung

Da sind wir sicherlich noch sehr, sehr weit weg von. Aber ich will nur darauf hinweisen: Erklärungsmethoden sind ganz aktuelle Forschung. Da... müssen wir halt auch noch ein bisschen weiter arbeiten.

Einen, der von einer anderen Richtung daran forscht, wie KI sich besser erklären kann, ist Niki Kilbertus. Wir haben ihn vorhin schon einmal kennen gelernt - er beschäftigt sich mit der Frage, ob man Kausalitäten, also Wenn-Dann-Zusammenhänge, in die Logik der KI-Systeme übersetzen kann. Aus den Daten werden dann kleine Grafiken, in denen Pfeile beschreiben, welche Ursache welche

Wirkung hervorruft.[11] Die Annahmen dahinter muss allerdings immer noch ein Mensch liefern.

> *Also, ich kann nicht aktuell ein System bauen, das einfach nur Daten aus der Welt aufnimmt und dann solche kausalen Zusammenhänge findet oder irgendwie gut ausdrücken kann. 04:05 Es funktioniert in sehr speziellen fällen, wo man weiß, dass wirklich alle Eventualitäten schon in den Daten vorkommen. Aber im allgemeinen müssen wir immer damit starten, das wir Annahmen treffen und dafür benötigt es menschliche Intelligenz.*

Zumal das mit dem Wenn und dem Dann gar nicht so einfach ist, auch für uns Menschen nicht. Rauchen verursacht Krebs - diese Aussage ist heute eine banale Feststellung, aber der Weg dorthin war mühsam. Niki Kilbertus erzählt die Anekdote von Ronald Fisher, einem britischen Statistiker, der in den 20er und 30er Jahren des vergangenen Jahrhunderts die Methoden der modernen Wissenschaft mit begründete. Allerdings war Fisher nicht nur ein brillianter Kopf, sondern auch ein überheblicher Kotzbrocken - und zudem starker Raucher.[12]

Als in den 50er Jahren immer mehr Mediziner einen Zusammenhang zwischen Rauchen und Lungenkrebs beobachten, führte er, unterstützt von der Tabakindustrie, einen bizarren Kampf zur Verteidigung des Nikotins, in der er alternative Erklärungen lieferte: Vielleicht war es ja so, dass Menschen mit einer genetischen Neigung zu Lungenkrebs besonders oft zur Zigarette griffen? Oder dass Rauchen

---

[11] Das Paper dazu: Kilbertus u.a., "Avoiding Discrimination through Causal Reasoning", eine journalistische Zusammenfassung hat Ariane Wetzel in Psychologie Heute veröffentlicht: Der faire Algorithmus.

[12] Schön beschrieben in diesem Artikel: https://priceonomics.com/why-the-father-of-modern-statistics-didnt-believe/ - Kleine Einschränkung: Die Quelle für den Artikel ist eine Marketing-Seite, keine Wissenschafts-Publikation, aber soweit ich die wissenschaftlichen Primärquellen gecheckt habe, ist der Artikel solide.

Wer es lieber etwas wissenschaftlicher hat: Hier die schöne Geschichte, als Fisher mit der Behauptung einer Kollegin(!) konfrontiert wurde, eine Tasse Milch mit Tee schmecke anders als Tee mit der gleichen Menge Milch - die sich überraschenderweise als richtig herausstellte. https://www.sciencehistory.org/distillations/ronald-fisher-a-bad-cup-of-tea-and-the-birth-of-modern-statistics

die Beschwerden von Lungenkranken linderte? Am Ende konnte die Wissenschaft Fishers Einwände widerlegen, aber die Geschichte unterstreicht einen Grundsatz: Korrelation ist keine Kausalität; dass die Anzahl von, sagen wir, Störchen und Babys in Deutschland zurückgeht, heißt noch nicht, dass das eine Ursache ist und das andere Wirkung. Wer ein Wenn mit einem Dann verbinden will, braucht mehr als einen statistischen Zusammenhang - er braucht ein Erklärungsmodell.

Algorithmen in kausale Zusammenhänge umwandeln, die entscheidenden Daten für die Nutzer kenntlich machen, und bei uns Menschen nachfragen - das sind viel versprechende Explainability-Ansätze. Wenn es den KI-Forscherinnen und Forschern so gelingt, dass die KI für uns Menschen keine Black Box mehr ist, gewinnen die Menschen und die Maschinen: Wir werden klüger, die KI wird besser.

Die verschlossene Blackbox, oder: Wie überzeugen wir Mark Zuckerberg?

So spannend und vielversprechend die Explainability-Forschung ist, sie hat einen Fehler: Sie setzt voraus, dass die Programmierer die Transparenz überhaupt wollen. Dort, wo Transparenz am wichtigsten wäre, weil KI-Algorithmen längst im Einsatz sind, wollen das die Firmen häufig nicht, die die KI gebaut haben. Auch wenn es wichtig wäre, mehr zu wissen und zu verstehen: KI-Riesen wie Google, Amazon, Microsoft und Facebook schützen ihre Geschäftsgeheimnisse.

NKB1 Journalist

Ich bin Nicolas Kayser-Bril, und ich bin ein Journalist für Algorithm Watch. Eine gemeinnützige Organisation in Berlin.

Der deutsch-französische Journalist arbeitet mit “Algorithm Watch” daran, zu erkunden und darzustellen, wie Algorithmen unser Bild von der Welt verändern können - und so nahmen er und einige Mitstreiterinnen einen Algorithmus ins Visier, der mitbestimmt, was

Millionen Menschen zu sehen bekommen: den Auswahl-Algorithmus von Instagram.[13]

NKB2 Instagram

Instagram ist das drittgrößte sozialen Netzwerk in Europa. Ungefähr 140 Millionen Menschen sind im auf Instagram aktiv, insbesondere die jüngere Generation, bei Leuten, die zwischen 18 und 25 sind, nutzen ungefähr hundert Prozent von denen Instagram. Also es ist riesig. Und das interessante ist, es gibt fast keine Recherche, keine Information, wie Instagram tatsächlich funktioniert.

Instagram gehört zu Mark Zuckerbergs Facebook-Imperium. Stimmt es, dass Facebook in etwa so sehr an Transparenz interessiert ist wie der Vatikan? Nicolas Kayser-Bril widerspricht dem nicht.

NKB3 Facebook sagt 1

Facebook sagt: das, was wir im Instagram zu sehen bekommen, spiegelt unsere Interessen. Und das ist alles.

Das reichte den Journalisten bei Algorithm Watch nicht - zumal sie da so einen Verdacht hatten: Ob Instagram-Fotos anderen Nutzer angezeigt wurden, konnte unter anderem von einem bestimmten Faktor stark beeinflusst werden. Der Tipp stammte von einer Influencerin:

NKB4 Bikini

...sie hatte das Gefühl, dass, wenn sie Bilder postete, dass ihre Bilder ihrer Follower erreichten, nur wenn sie im Bikini postete - und wir wollten diese Statements verifizieren

Hatte der Instagram-Algorithmus gelernt, nackte Haut zu bevorzugen? Nicolas und seine Mitstreiterinnen sichteten Patentschriften von Facebook, in denen beschrieben wurde, wie der Algorithmus

[13] Ergebnisse der Recherche hier: "Instagram-Algorithmus: Wer gesehen werden will, muss Haut zeigen"

vorherzusagen versucht, welche Instagram-Fotos Nutzerreaktionen auslösen.

NKB5 Patente

Und um das zu machen, analysieren Sie die, ob es Männer oder Frauen im Bild sind oder wie viel Haut zu sehen ist.

Das allerdings ist allenfalls ein Indiz, kein Beweis. Einblicke in die Black Box des Instagram-Algorithmus bekamen Nicolas und sein Team dadurch nicht. Also bemühten sie sich, die Black Box so genau wie möglich bei der Arbeit zu beobachten - um dadurch Rückschlüsse ziehen zu können. Sie entwickelten ein Browser-Add-On, ein kleines Programm, das bei knapp 30 Freiwilligen ihre Instagram-Zugriffe analysierte.

NKB6 Newsfeed

Und damit hatten wir einen Blick auf ihren eigenen Instagram Newsfeeds. Und diese Daten haben wir dann gespeichert und analysiert, und damit konnten wir zeigen, dass Bilder von Influencerinnen, die Haut zeigten, tatsächlich mehr Wahrscheinlichkeit hatten, in der Newsfeed von den Nutzern zu sein.

Und das war statistisch signifikant, oder anders gesagt: es konnte kein Zufall sein. Ob die Vorliebe der untersuchten Feeds für nackte Haut nur ein Kluger-Hans-artiges Nebenprodukt der verwendeten Trainingsdaten war, oder von Facebook gewollt, war so nicht herauszufinden. Schon gar nicht von Facebook.

Die dürre Reaktion von Facebook auf die Recherche-Ergebnisse: Stimmt doch gar nicht. Fehlerhafte Analyse, viel zu wenige Einzelfälle, Punkt. Der Instagram-Algorithmus zeige einfach nur die je nach Nutzer interessantesten Bilder.

Was aber “interessant” heißt, und was den Ausschlag gab - das hätte man vielleicht mit Explainability-Ansätzen wie dem vorher erwähnten

LIME-Algorithmus gut darstellen können, aber von Facebook gab es dazu kein Sterbenswörtchen.[14]

Müssen wir also Facebook und all die anderen Riesen dazu zwingen, Erklär-Mechanismen in ihre Algorithmen einzubauen? Könnte man machen, sagt Nicolas Kayser-Bril, aber das setze ein Vertrauen in Facebook voraus, das er dem Konzern nicht entgegenbringt.

> **NKB9 Algopolizei**
>
> *Man könnte Gesetze schreiben, damit Facebook und Co Explainability einsetzen müssten. Das Problem ist: sie könnten ihre Systemen so bauen, damit sie nicht die ganze Wahrheit wiedergeben würden. Und in Deutschland haben wir sehr viel Erfahrung damit, mit Volkswagen und die anderen Autohersteller, die solche Lügenmechanismen gebaut haben.*
>
> *Und deswegen ist die einzige Lösung wirklich der physischen Zugang zu den Servern von Facebook zu bekommen, und das sollte ein bisschen wie Kontrollen in Restaurants passieren, dass eine Algorithmus-Polizei wirklich direkt in diese Data Center gehen würde und ihre eigenen Einblicke in den Code? bekommen könnte.*

Möglicherweise ist das der Grund, warum Explainability gar nicht so eine große Rolle spielen wird: Wenn Erklärungsmechanismen in die KI eingebaut werden, müssen wir auch dem trauen können, der sie eingebaut hat; siehe VW…..

Die Grenzen der Erklärbarkeit

Ein unabhängiger Algorithmen-TÜV - diesem Gedanken kann der Maschinenlern-Experte Kristian Kersting von der TU Darmstadt etwas abgewinnen. Auch wenn er tendenziell staatlicher Regulierung seines Forschungsgebiets skeptisch gegenüber steht.

---

[14] Algorithmwatch recherchiert weiter - wenn Sie sich beteiligen wollen, können auch Sie sich das angesprochene Browser-Plugin installieren und Ihre Daten spenden: https://algorithmwatch.org/instagram-algorithmus/ - das setzt natürlich voraus, dass Sie Algorithmwatch vertrauen. Ich tu's.

Kersting: Also ich finde schon: so, wie in der Medizin ... Medikamente geprüft werden müssen, glaube ich: Algorithmen, die wichtig sind und die auf die Allgemeinheit losgelassen werden, sollten auch noch einmal speziell irgendwie geprüft werden. Und da können wir uns sicherlich von der Medizin auch viel abgucken. Aber ich finde es eben auch wichtig zu sagen: es zeigt uns, dass wir nicht immer ganz richtig sind und perfekt sind. Und die Maschine kann uns auch helfen, unter Umständen irgendwann genau das festzustellen.

Bei der Analyse der Arbeit einer KI wird immer wieder herauskommen: Dass die Maschine von den Menschen mit den Entscheidungen zugleich ihre Vorurteile erlernt hat.

Umgekehrt haben Menschen eine fatale Neigung, die Entscheidungen von Algorithmen als wissenschaftlich fundiert zu akzeptieren. Die Algorithmen-Forscherin Katharina Zweig in Kaiserslautern[15] hat das erlebt. Sie erzählt von einem Forschungsprojekt, an dem sie beteiligt war. Es ging um ein Empfehlungssystem für Kinofilme, das ziemlich merkwürdige Empfehlungen gab - aufgrund eines mathematischen Fehlers empfahl es Fans der Liebeskomödie Pretty Woman auch das Science-Fiction-Spektakel Star Wars. Klar, sagten sich einige Forscher daraufhin, wer großes Hollywood-Kino mag, bekommt mehr davon empfohlen, da hat die KI schon Recht. Und Pretty Woman gehört wie Star Wars zum großen Hollywood -Kino Die Menschen erfanden sich also eine Geschichte, um sich die (falschen) Entscheidungen der KI zu erklären. Ohne zu wissen, ob das der wirkliche Auslöser war.

Zweig Leichte Beute

Und das hat mir schon gezeigt, dass wir Menschen sehr schlecht sind, darin abweichende Ergebnisse, die eigentlich völliger Unsinn sind, auch als solche zu erkennen, solange es irgendwie noch ne Geschichte gibt, die wir uns dazu erzählen können.

---

[15] Sie forscht unter anderem an ihrem "Algorithmic Accountability Lab" mit ihrem Team darüber, ob und wie Algorithmen Bevölkerungsgruppen benachteiligen, und ist Mitbegründerin von "Algorithm Watch".

JE: Ja, aber das ist doch fürchterlich gefährlich, wenn wir Menschen so angelegt sind... dass wir uns immer irgendwie rückwirkende Geschichte erfinden dazu, irgendwie eine Story. sein. Dann sind wir leichte Beute für die Algorithmen.

Zweig 29:53 Ja, das sind wir tatsächlich! Und deswegen braucht es auch einen Prozess, der sicherstellt, dass das nicht passiert.

...sagt Katharina Zweig, die übrigens vor Jahren Algorithm Watch mitbegründet hat, die Organisation, für die der Journalist Nicolas Kayser-Bril arbeitet. Er skizziert, wie so ein Prozess aussehen könnte:

NKB8 Audit

Im Idealfall würden Behörden Zugang zu Trainingsdaten haben, aber auch Zugang zu den Produktions-Servern, also zu eigenen Servern von Facebook. Und damit könnten sie wirklich testen, wie der Algorithmus funktioniert. Trainingsdaten alleine bringen leider nicht sehr viel.

Einblick in die Black Box - Transparenz darüber, wie Algorithmen arbeiten und entscheiden, entsteht nicht von selbst. Wir, die Gesellschaft, müssen die entsprechenden Regeln erst schaffen. Denn dass beispielsweise Facebook von selbst für klare, erschöpfende und richtige Erklärungen sorgt - das Vertrauen darauf dürfte eher eingeschränkt sein.