Explainable AI (ex-AI)

Andreas Holzinger

"One of the first things taught in introductory statistics textbooks is that correlation is not causation. It is also one of the first things forgotten."
Thomas Sowell, Stanford (1930–)

Motivation

„Artificial Intelligence" (AI) besitzt eine lange Tradition in der Informatik und erlebte in den letzten sechs Jahrzehnten viele Höhen und Tiefen. Das Feld hat in letzter Zeit vor allem aufgrund großer praktischer Erfolge statistischer und probabilistischer Ansätze in Machine Learning (ML) enormes Interesse geweckt. Das große Ziel ist nach wie vor, Software zu entwickeln, die in der Lage ist, selbstständig aus Erfahrungen zu lernen und Voraussagen zu treffen. Um ein Niveau an praktisch nutzbarer AI zu erreichen ist es notwendig: (1) aus hochdimensionalen Datenmengen zu lernen, (2) daraus Wissen zu extrahieren, (3) dieses zu verallgemeinern, (4) dabei aber den „Fluch der Dimensionalität" in den Griff zu bekommen, und schließlich (5) die den Daten zugrundeliegenden **Erklärungsfaktoren** zu verstehen. Letzteres impliziert allerdings die wahrscheinlich größte Herausforderung moderner AI: Daten im **Kontext** einer Anwendungsdomäne zu **verstehen.**

Intelligente selbstlernende Algorithmen bzw. Systeme sollten in der Lage sein, sowohl den sprachlichen Kontext (verbal) als auch den situativen Kontext (nonverbal) zu verstehen und daraus nicht nur formal, sondern auch kausal die korrekten Schlüsse zu ziehen – ein langersehntes, aber noch weit entferntes Ziel der AI.

Derzeit sind AI-Anwendungen in unserem täglichen Leben sehr erfolgreich (Autonomes Fahren, Sprachverstehen, Empfehlungssysteme, usw.). Für Menschen ist es jedoch sehr schwierig nachzuvollziehen, **wie** diese Algorithmen zu einer Entscheidung gelangen. Ansätze, wie z. B. „Deep Learning" [15] sind letztlich sogenannte „Black-Box"-Modelle. Selbst wenn wir die zugrunde liegenden mathematischen Prinzipien verstehen, fehlt solchen Modellen eine explizite deklarative Wissensrepräsentation. Interessanterweise hatten frühe AI-Lösungen (damals Expertensysteme genannt) von Anfang an das Ziel, Lösungen nachvollziehbar, verstehbar und damit erklärbar zu machen, was in eng begrenzten Domänen auch möglich war [11].

Natürlich benötigt man nicht für alles und jederzeit Erklärungen. Eigentlich ist ja genau das Gegenteil der Fall und deswegen ist AI mit ihren statistischen Lernmethoden derzeit so erfolgreich: Abstrakte Algorithmen finden in großen, komplexen und hochdimensionalen Datenmengen Muster, die kein Mensch jemals entdecken könnte. Das ist gut. Allerdings gibt es bestimmte Domänen und bestimmte Situationen, in denen eine nachvollziehbare Erklärung notwendig ist. Insbesondere gilt dies in problematischen Situationen menschlicher Entscheidungsfindung. Hier kann eine Erklärungskomponente dazu beitragen,

DOI 10.1007/s00287-018-1102-5

Andreas Holzinger
Medizinische Universität Graz,
Institut für Medizinische Informatik,
Statistik & Dokumenation, Holzinger Group HCI-KDD,
und Technische Universität Graz,
Fakultät für Informatik & Biomedizinische Technik,
Institute for Interactive Systems and Data Science,
Auenbruggerplatz 2/V, 8036 Graz, Österreich
E-Mail: a.holzinger@hci-kdd.org

Alle „Aktuellen Schlagwörter" seit 1988 finden Sie unter:
http://www.is.informatik.uni-wuerzburg.de/as

Published online: 03 April 2018

Zusammenfassung

„Explainable AI“ ist kein neues Gebiet. Vielmehr ist das Problem der Erklärbarkeit so alt wie die AI selbst, ja vielmehr das Resultat ihrer selbst. Während regelbasierte Lösungen der frühen AI nachvollziehbare „Glass-Box“-Ansätze darstellten, lag deren Schwäche im Umgang mit Unsicherheiten der realen Welt. Durch die Einführung probabilistischer Modellierung und statistischer Lernmethoden wurden die Anwendungen zunehmend erfolgreicher – aber immer komplexer und opak. Beispielsweise werden Wörter natürlicher Sprache auf hochdimensionale Vektoren abgebildet und dadurch für Menschen nicht mehr verstehbar. In Zukunft werden kontextadaptive Verfahren notwendig werden, die eine Verknüpfung zwischen statistischen Lernmethoden und großen Wissensrepräsentationen (Ontologien) herstellen und Nachvollziehbarkeit, Verständlichkeit und Erklärbarkeit erlauben – dem Ziel von „explainable AI“.

den menschlichen Entscheidern zumindest eine Chance auf Überprüfung der Plausibilität eines Ergebnisses zu ermöglichen. Ein Beispiel sind medizinische Entscheidungsunterstützungssysteme. Hier sind Lösungen hilfreich, die es ermöglichen, Entscheidungen nachvollziehbar transparent, verständlich und erklärbar zu machen. Gerade in sicherheitsrelevanten Domänen stellt sich nämlich zwangsläufig die Frage: „Können wir unseren Ergebnissen vertrauen?“ [4].

Hier ist „explainable AI“ nicht nur nützlich und notwendig, sondern stellt überdies eine Riesenchance für AI-Lösungen generell dar, weil dadurch die vorgeworfene Undurchsichtigkeit der AI vermindert und notwendiges Vertrauen aufgebaut werden kann. Genau dies kann die Akzeptanz bei zukünftigen Benutzern nachhaltig fördern. Ein weiterer und wichtiger werdender Bereich ist der juristische. Hier drängt die Zeit, dass die Informatikforschung Lösungen findet insofern, als die neue Europäische Datenschutzgrundverordnung (DSGVO, vgl. auch mit ISO/IEC 27001) ein „Recht auf Erklärung“ vorsieht. Dies bedeutet keinesfalls, alles und immerzu in Echtzeit erklären zu müssen; jedoch auf Antrag einer Person, eine Erklärung für eine bestimmte Entscheidung oder eine Risikobewertung nachvollziehbar und erklärbar darzustellen.

All die genannten Umstände machen das aktuelle Schlagwort „explainable AI“ zu einem Thema, das weltweit, sowohl in Wissenschaft als auch in Wirtschaft enorm an Bedeutung zunimmt und verstärkt zur Diskussion von Transparenz, Vertrauen, Interpretierbarkeit, Nachvollziehbarkeit und Erklärbarkeit aber auch von ethischen Aspekten von AI zwingt.

Begrifflichkeiten: Verstehbar? Verständlich? Erklärbar?

Der Begriff AI selbst ist eigentlich ein unglücklicher, ist doch gerade das Phänomen natürlicher Intelligenz sehr schwer zu definieren und von einer Fülle verschiedener Faktoren abhängig; daher beschränken wir uns hier nur auf explizit Relevantes für das Schlagwort „explainable AI“.

Verstehen ist nicht nur erkennen, wahrnehmen und wiedergeben (Reizreaktion auf physiologischer Ebene), und auch nicht nur das inhaltliche Begreifen und bloße Wiedergeben eines Sachverhalts, sondern die intellektuelle Erfassung des Zusammenhangs (Kontext), in dem dieser Sachverhalt steht. Verstehen ist vielmehr die Brücke zwischen Wahrnehmen und Entscheiden. Von der Erfassung des Kontextes, zweifelsfrei ein wichtiger Indikator für Intelligenz schlechthin, ist die derzeitige AI aber noch meilenweit entfernt. Dagegen sind Menschen aber sehr gut in der Lage, den Kontext instantan zu erfassen und bereits aus sehr wenigen Datenpunkten sehr gute Generalisierungen vorzunehmen [6].

Erklären (Interpretieren) bedeutet darüber hinaus, die Ursachen eines beobachteten Sachverhaltes durch eine sprachliche Darlegung seiner logischen und kausalen Zusammenhänge verständlich zu machen. In der Wissenschaftstheorie gilt gemäß dem hypothetisch-deduktiven Modell nach Karl Popper eine kausale Erklärung als Fundament jeder Wissenschaft, um Sachverhalte aus Gesetzen und Bedingungen deduktiv abzuleiten. Kausalität und kausales Schlussfolgern ist daher ein extrem wichtiges Gebiet für „explainable AI“ [10].

Verstehen und Erklären sind Voraussetzungen für Nachvollziehbarkeit. Die Frage die sich uns nun stellt ist: *„Was ist für den Menschen überhaupt verstehbar bzw. verständlich?“*

Direkt verständlich und damit auch erklärbar, interpretierbar und nachvollziehbar für Menschen

sind Daten bzw. Objekte $\leq \mathbb{R}^3$, z. B. Bilder (Matrix aus Pixeln, Glyphen, Korrelationsgraphen, 2D/3D-Projektionen usw.) oder Text (Sequenzen natürlicher Sprache). Menschen können Bilder bzw. Wörter physiologisch perzipieren, die extrahierte Information entsprechend kognitiv mit Bezug auf ihr subjektives Vorwissen interpretieren (verstehen) und in den jeweiligen, eigenen kognitiven Wissensraum integrieren. Streng genommen muss hier zwischen Bildverstehen, Textverstehen und Sprachverstehen unterschieden werden. Für weiterführende Information wird hier der Kürze wegen auf die Kognitionsforschung verwiesen.

Nicht direkt verständlich und damit auch nicht erklärbar, interpretierbar und nachvollziehbar sind abstrakte Vektorräume in $> \mathbb{R}^3$(z. B. „word-embeddings“) oder undokumentierte, d. h. noch unbekannte Eingangsmerkmale (z. B. Textsequenzen mit unbekannten Wörtern oder unbekannten Symbolen). Ein Beispiel soll dies verdeutlichen: Beim sogenannten „Word-Embedding“ [9] werden Wörter und/oder Phrasen jeweils Vektoren zugeordnet. Konzeptionell ist dies eine mathematische Einbettung von einem Raum mit einer Dimension pro Wort in einen kontinuierlichen Vektorraum mit geringerer Dimension. Methoden, um ein solches „Mapping“ zu generieren, umfassen z. B. neuronale Netze und probabilistische Modelle mit einer expliziten Repräsentation in Bezug auf den Kontext, in dem Wörter erscheinen.

Post-Hoc- und Ante-Hoc-Erklärungsmodelle

A) Post-Hoc Erklärungsansätze

Post-Hoc (lat.) = nach-diesem (Ereignis), d. h. solche Ansätze liefern eine Erklärung für eine spezifische Lösung, erklären also nicht das gesamte Modell. Der Kürze wegen wird jeweils nur ein Beispiel etwas genauer vorgestellt.

Bei **BETA (Black Box Explanations through Transparent Approximations)** [7] handelt es sich um ein agnostisches Modell zur Erklärung des Verhaltens eines (beliebigen) Black-Box-Klassifikators (also einer Funktion, die einen Merkmalsraum auf eine Menge von Klassen abbildet) durch gleichzeitige Optimierung auf Genauigkeit des ursprünglichen Modells und einer Interpretierbarkeit der Erklärung für einen Menschen. Interpretierbarkeit und Genauigkeit gleichzeitig sind schwierig zu erreichen. Die Benutzer werden interaktiv in das Modell eingebunden und können so die sie interessierenden Bereiche von Black-Box-Modellen erkunden.

Das **LRP (Layer-Wise Relevance Propagation)**-Verfahren [1] stellt eine weitere allgemeine Lösung zum Verstehen von Klassifikationsentscheidungen durch pixelweise Zerlegung von nichtlinearen Klassifikatoren dar. Stark vereinfacht erlaubt LRP, die „Denkprozesse“ von neuronalen Netzen rückwärts ablaufen zu lassen. Dabei wird nachvollziehbar, welcher Input welchen Einfluss auf das jeweilige Ergebnis hatte, z. B. im Einzelfall, wie das neuronale Netz zu einer medizinischen Diagnose oder einer Risikobewertung gekommen ist. Werden genetische Daten in ein solches Netz eingegeben, kann nicht nur analysiert werden, mit welcher Wahrscheinlichkeit ein Patient eine bestimmte genetische Erkrankung hat, sondern anhand **welcher** Merkmale diese Entscheidung getroffen wurde. Ein solcher Ansatz ist definitiv ein Schritt in Richtung personalisierter Medizin. In Zukunft kann mit solchen Ansätzen eine individuelle, genau auf den Patienten „zugeschnittene“ Krebstherapie erfolgen.

LIME (Local Interpretable Model-Agnostic Explanations) [12] stellt ein agnostisches Modell dar, worin $x \in \mathbb{R}^d$ die ursprüngliche Repräsentation einer zu erklärenden Instanz und $x' \in \mathbb{R}^{d'}$ einen Vektor für die zu interpretierende Repräsentation darstellt. Beispielsweise kann x' ein Merkmalsvektor sein, der Wörter x' eines „bag-of-words“-Ansatzes enthält.

Das Ziel ist, ein interpretierbares Modell zu finden, welches lokal vertrauenswürdig zum Klassifikator passt, d. h.:

$$g\colon \mathbb{R}^{d'} \rightarrow \mathbb{R}, g \in G$$

wobei G die Klasse der interpretierbaren Modelle darstellt, z. B. lineare Modelle, Entscheidungsbäume, Regellisten, usw.; ein gegebenes Modell $g \in G$ kann somit zur Erklärung für einen menschlichen Experten in $\mathbb{R}^2$ dargestellt werden. LIME arbeitet mit jeder Instanz *separat,* diese werden permutiert und ein Ähnlichkeitsmaß zu den ursprünglichen Instanzen berechnet. Nun lässt man das komplexe Modell Vorhersagen für jede dieser permutierten Instanzen machen. Der Einfluss der Änderungen auf die Vorhersagen kann für jede Instanz nachvollzogen werden. so kann beispielsweise ein Arzt

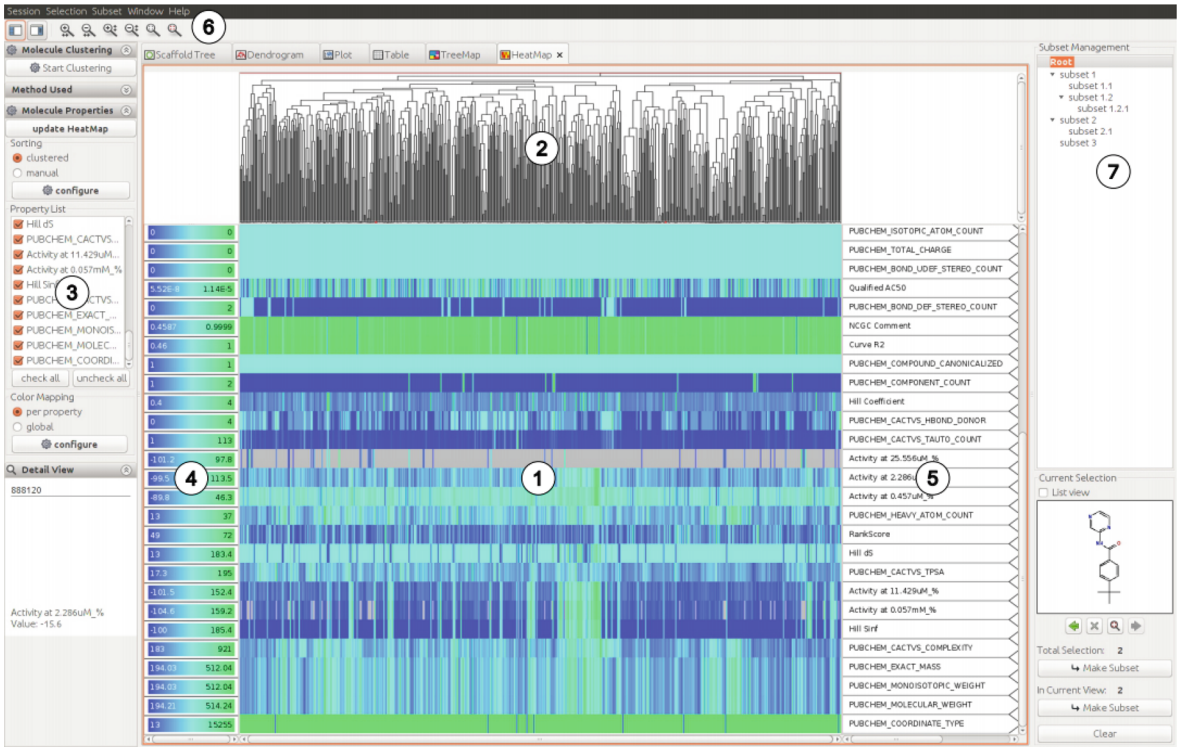

Abb. 1 Heatmap zur Datenvisualisierung von Molekülen vs. Moleküleigenschaften. Die Sicht ist eine interaktive „Table View“* (1), *d. h. die Spalten sind Moleküle, die durch einen hierarchischen Clusteralgorithmus nach Ähnlichkeit gruppiert sind. Das Dendrogramm oben* (2) *zeigt die Gruppierung und so kann der Benutzer entscheiden, welche Moleküle eine Gruppe bilden. Die Zeilen sind Merkmale der Moleküle, also z. B. Wirksamkeiten, chemische und andere Eigenschaften, die sich numerisch messen lassen. Die Werte pro Molekül werden farbcodiert; so sieht man pro Merkmal wie sich die Gruppen unterscheiden. Die Exploration passiert durch Interaktion und Drill Down, u. a. um Struktur- bzw. Verbindungserklärungen zu bekommen;* (3) *zeigt Einstellmöglichkeiten, um die Heatmap zu konfigurieren; zwischen der linken Sidebar und der Heatmap findet man die Legende für die Werte* (4), *rechts die Namen der Moleküle* (4) *und* (5)*; (6) die Steuerelemente und* (7) *zusätzlich eine tree-map. Man kann zusätzlich die Parameter der Clusterisierung ändern, um andere Gruppierungen zu erhalten [14].

überprüfen, ob Ergebnisse realistisch sind (zum Beispiel zur Voraussage von Wiederzuweisungen in ein Krankenhaus).

B) Ante-Hoc-Erklärungsansätze

Ante-hoc-Methoden sind systemimmanent interpretierbar, also von Natur aus transparent (Glass-Box), ähnlich wie beim „interactive Machine Learning“ (iML)-Modell [5]. Viele Ante-hoc-Ansätze erscheinen besonders neuartig, aber gerade diese Ansätze haben eine lange Tradition und wurden in Expertensystemen seit Beginn der AI eingesetzt, insbesondere Entscheidungsbäume, lineare Regression und Random Forests, um drei zu nennen.

Ein wichtiger Aspekt ist die quantitative Beurteilung der **Qualität der Erklärungen.** Bei Ante-Hoc-Systemen, z. B. kann oft ein relevanter Messparameter definiert werden. In Fuzzy-Systemen ist das besonders einfach, da die Interpretierbarkeit mit der Anzahl der Regeln bzw. der Regelbedingungen zusammenfällt; dies gelang besonders gut bei so genannten „Intelligent Tutoring Systems“ (ITS) [13].

Generalisierte Additive Modelle (GAMs) sind hinsichtlich Verständlichkeit sehr nützlich, solange niedrigdimensionale und damit menschlich verstehbare Terme (also für Menschen lesbare Texte wie in Abb. 1 links) berücksichtigt werden. Ein Ansatz wurde von Caruana et al. [2] präsentiert: Hocheffiziente GAMs wurden mit paarweisen Interaktionen auf medizinische Problemstellungen angewandt. Solche Modelle sind für medizinische Experten verständlich und erlauben die Entdeckung von Mustern in den Daten – die ansonsten verborgen geblieben wären. Für „explainable AI“ ist das interessante dabei, dass der Einfluss *jedes* Merkmals

if hemiplegia **and** age > 60 **then** *stroke risk* 58.9% (53.8%–63.8%)
else if cerebrovascular disorder **then** *stroke risk* 47.8% (44.8%–50.7%)
else if transient ischaemic attack **then** *stroke risk* 23.8% (19.5%–28.4%)
else if occlusion and stenosis of carotid artery without infarction **then** *stroke risk* 15.8% (12.2%–19.6%)
else if altered state of consciousness **and** age > 60 **then** *stroke risk* 16.0% (12.2%–20.2%)
else if age ≤ 70 **then** *stroke risk* 4.6% (3.9%–5.4%)
else *stroke risk* 8.7% (7.9%–9.6%)

Abb. 2 Bestimmung des Schlaganfallrisikos nach der Diagnose von Vorhofflimmern aus der Anamnese des Patienten. Das gegebene Risiko ist der Mittelwert der nachfolgenden Verteilung (in Klammern jeweils das 95 % Intervall). Solche Bayesian Rule Lists können von Ärzten nachvollzogen, auf Plausibilität geprüft und, was besonders wertvoll ist, es können Schwächen des jeweiligen Vorhersagemodells erkannt werden (Abbildung stammt aus [8]).

auf das Ergebnis z. B. durch Heatmaps (siehe Abb. 1) visualisiert werden kann. So werden die Ergebnisse für menschliche Experten nachvollziehbar, was z. B. bei der Modellierung von Erkrankungswahrscheinlichkeiten in klinischen Studien hilfreich ist, oder zur Identifikation von Risikofaktoren bei Kreditinstituten.

Hybride Systeme werden oft in der Medizin verwendet, wo man nicht nur mit Bilddaten und „–omics"-Daten (z. B. genetischen Daten), sondern auch mit komplexen Mengen von Text zu tun hat, ist die Kombination von traditionellen logikbasierten Systemen, der Einbindung vorhandener großer Wissensbasen zusammen mit statistischen und probabilistischen Ansätzen (z. B. Deep Learning) sehr vielversprechend. Die Medizin ist ein Prototyp für nichtmonotones Schließen, wo man Schlüsse ziehen und Entscheidungen unter großer Unsicherheit treffen muss. Zudem ist dieser Bereich durch unvollständige Informationen gekennzeichnet. Gerade aber aufgrund der hohen semantischen Mehrdeutigkeit wurden in dieser Domäne schon sehr früh große Mengen an Wissensbasen manuell erstellt, z. B. die Gene Ontology (GO), die es erlauben auf eindeutige Begrifflichkeiten zurückzugreifen. Wertvolle Erkenntnisse für Entscheidungen sind oft in der Verknüpfung vorhandener Daten verborgen.

Um hier erklärbare Strukturen zu gewinnen, müssen Daten aus unterschiedlichsten Quellen fusioniert, verknüpft und validiert werden – gerade hier kann ein Domänenexperte entscheidend zu Erklärungskomponenten beitragen und es kann nicht oft genug betont werden, wie wichtig im Bereich des maschinellen Lernens das Domänenwissen ist [3].

Ein Beispiel ist der Ansatz von Letham et al. [8], der es erlaubt, Vorhersagemodelle zu erstellen, die nicht nur genau, sondern auch für menschliche Experten interpretierbar sind. Solche Modelle bestehen aus Entscheidungslisten mit einer Reihe von WENN-DANN-Aussagen (z. B. WENN hoher Blutdruck, DANN Schlaganfall). So kann ein hochdimensionaler, multivariater Merkmalsraum in einen niedrigdimensionalen und somit menschlich interpretierbaren Entscheidungsraum transferiert (diskretisiert) werden.

In [8] wird dazu ein generatives Modell namentlich „Bayesian Rule Lists (BRL)" verwendet, welches eine posterior-Verteilung über mögliche Entscheidungslisten erlaubt (Abb. 2). Experimente zeigten, dass Bayes'sche Regellisten eine Vorhersagegenauigkeit aufweisen, die mit den besten aktuellen Algorithmen, z. B. „Support Vector Machines" und „Classification and Regression Trees" (CART), vergleichbar ist.

Diese Methode wird ebenfalls durch die zunehmende Notwendigkeit von Erklärungskomponenten in der personalisierten Medizin motiviert und kann verwendet werden, um wesentlich genauere und interpretierbare medizinische Scoringsysteme zu erzeugen. Ein solches Scoringsystem zur klinischen Risikoanalyse ist beispielsweise $CHADS_2$ (die Abkürzung steht für Congestive heart failure, Hypertension, Age, Diabetes, Prior Stroke); und in [8] wurde gezeigt, dass deren Modell ebenso interpretierbar wie $CHADS_2$ ist – aber wesentlich genauer.

Chancen, offene Fragen und zukünftige Herausforderungen

Die große Chance von „explainable AI" ist nicht nur „Black Boxes" transparent zu machen und damit Vertrauen in AI zu fördern, sondern vor allem ein tieferes Verständnis für vorher unbekannte Zusammenhänge zu fördern. Man denke nur an die

enorme Hilfestellung, die Ärzte aus der Kombination von menschlicher Intelligenz und AI (bspw. während einer Diagnosefindung) beziehen können: Menschen zeigen in niedrigdimensionalen Problemstellungen sehr gute Intuition, können durch ihre Alltagsintelligenz erstaunlich gut aus wenigen Daten generalisieren und Zusammenhänge erkennen.

So könnten sie beispielsweise AI auf „interessante" Daten ansetzen und interaktiv hinterfragen. Umgekehrt können maschinell aus hochdimensionalen Datenräumen erhaltene Resultate, die kein Mensch je hätte finden können, nachvollzogen und auf Plausibilität geprüft werden. Vielleicht der wichtigste Beitrag von „explainable AI" ist es, aufzuklären, was Ursache ist und was Wirkung (und welches nur Korrelation) – um zu vermeiden, dass man fälschlich Artefakte und Surrogate miteinbezieht. Dies ist in vielen Anwendungsdomänen wünschenswert, in sicherheitskritischen Bereichen sogar zwingend erforderlich.

Die große Chance für die Zukunft besteht aus einer Verknüpfung verschiedener bereits bewährter Ansätze, z. B. logikbasierte Ontologien mit probabilistischem, maschinellem Lernen mit einem (oder mehreren bzw. sogar vielen) human-in-the-loop zu einem hybriden Multiagenten-Interaktionsmodell zu fusionieren, in der AI als eine Art „Servolenkung fürs Gehirn" unterstützend verwendet wird. Dies würde nicht nur eine Erweiterung (Augmentation) menschlicher Intelligenz mit maschineller Intelligenz bedeuten, sondern auch umgekehrt eine Erweiterung der künstlichen Intelligenz durch menschliche Intuition.

Ein solcher Ansatz ist mittelfristig vermutlich die einzig erfolgversprechende Möglichkeit, Systeme zu entwickeln, die in der Lage sind, **kontextuelle Erklärungsmodelle für Klassen realer Phänomene** zu konstruieren.

Danksagung

Der Autor dankt den anonymen Begutachtern für ihre freundlichen Kommentare und wertvollen Hinweise.

Funding. Open access funding provided by Medical University of Graz.

Literatur

1. Bach S, Binder A, Montavon G, Klauschen F, Müller KR, Samek W (2015) On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. Plos One 10(7):e0130140, doi:10.1371/journal.pone.0130140
2. Caruana R, Lou Y, Gehrke J, Koch P, Sturm M, Elhadad N (2015) Intelligible Models for Healthcare: Predicting Pneumonia Risk and Hospital 30-Day Readmission. In: 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '15), Sydney, 2015. ACM, doi:10.1145/2783258.2788613, pp 1721–1730
3. Girardi D, Küng J, Holzinger A (2015) A Domain-Expert Centered Process Model for Knowledge Discovery in Medical Research: Putting the Expert-in-the-Loop. In: Guo Y, Friston K, Aldo F, Hill S, Peng H (eds) Brain Informatics and Health, Lecture Notes in Computer Science LNCS 9250. Springer, Cham Heidelberg Berlin London Dordrecht New York, pp 389–398
4. Holzinger K, Mak K, Kieseberg P, Holzinger A (2018) Can we trust Machine Learning Results? Artificial Intelligence in safety-critical Decision Support. ERCIM News 112(1):42–43
5. Holzinger A, Plass M, Holzinger K, Crisan GC, Pintea CM, Palade V (2017) A Glass-Box Interactive Machine Learning Approach for Solving NP-hard Problems With The Human-in-the-Loop. arXiv:1708.01104
6. Lake BM, Ullman TD, Tenenbaum JB, Gershman SJ (2017) Building machines that learn and think like people. Behav Brain Sci 40(e253), doi:10.1017/S0140525X16001837
7. Lakkaraju H, Kamar E, Caruana R, Leskovec J (2017) Interpretable and Explorable Approximations of Black Box Models. arXiv:1707.01154
8. Letham B, Rudin C, McCormick TH, Madigan D (2015) Interpretable classifiers using rules and Bayesian analysis: Building a better stroke prediction model. Ann Appl Stat 9(3):1350–1371
9. Mikolov T, Sutskever I, Chen K, Corrado GS, Dean J (2013) Distributed Representations of Words and Phrases and Their Compositionality. In: Burges CJC, Buttou L, Welling M, Ghahramani Z, Weinberger KQ (eds) Advances in Neural Information Processing Systems 26 (NIPS 2013). pp 3111–3119
10. Peters J, Janzing D, Schölkopf B (2017) Elements of Causal Inference: Foundations and Learning Algorithms. MIT Press, Cambridge (MA)
11. Puppe F (1993) Einführung in Expertensysteme. Springer, Heidelberg
12. Ribeiro MT, Singh S, Guestrin C (2016) Why Should I Trust You? Explaining the Predictions of Any Classifier. In: 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp 1135–1144
13. Schewe S, Quak T, Reinhardt B, Puppe F (1996) Evaluation of a Knowledge-Based Tutorial Program in Rheumatology – A Part of a Mandatory Course in Internal Medicine. In: Frasson C, Gauthier G, Lesgold A (eds) International Conference on Intelligent Tutoring Systems (ITS 1996), LNCS 1986. Springer, Heidelberg, pp 531–539
14. Sturm W, Schaefer T, Schreck T, Holzinger A, Ullrich T (2015) Extending the Scaffold Hunter Visualization Toolkit with Interactive Heatmaps. In: Borgo R, Turkay C (eds) EG UK Computer Graphics & Visual Computing CGVC 2015, University College London (UCL), Euro Graphics (EG), pp 77–84
15. Wick C (2017) Deep Learning. Informatik-Spektrum 40(1):103–107