Verbinden von Natürlicher und Künstlicher Intelligenz: eine experimentelle Testumgebung für Explainable AI (xAI)

Andreas Holzinger · Heimo Müller

Eingegangen: 14. Oktober 2019 / Angenommen: 18. Januar 2020 / Online publiziert: 30. Januar 2020
© Der/die Autor(en) 2020

Zusammenfassung Künstliche Intelligenz (KI) folgt dem Begriff der menschlichen Intelligenz, der leider kein klar definierter Begriff ist. Die gebräuchlichste Definition, wie sie in der Kognitionswissenschaft als mentale Fähigkeit gegeben ist, enthält unter anderem die Fähigkeit, abstrakt, logisch und schlussfolgernd zu denken und gegebene Probleme der realen Welt zu lösen. Ein aktuelles Thema in der KI ist es, herauszufinden, ob und inwieweit Algorithmen in der Lage sind, solches abstraktes Denken und Schlussfolgern ähnlich wie Menschen zu erlernen – oder ob das Lernergebnis auf rein statistischer Korrelation beruht. In diesem Beitrag stellen wir eine von uns entwickelte frei verfügbare, universelle und erweiterbare experimentelle Testumgebung vor. Diese „Kandinsky Patterns“ (<https://human-centered.ai/project/kandinsky-patterns>, <https://www.youtube.com/watch?v=UuiV0icAIRs>), benannt nach dem russischen Maler und Kunsttheoretiker Wassily Kandinsky (1866–1944), stellen eine Art „Schweizer Messer“ zum Studium der genannten Problemstellungen dar. Das Gebiet, dass diese Problemstellungen behandelt wird „explainable AI“ (xAI) genannt. Erklärbarkeit/Interpretierbarkeit hat das Ziel, menschlichen Experten zu ermöglichen, zugrundeliegende Erklärungsfaktoren – die Kausalität – zu verstehen, also *warum* eine KI-Entscheidung getroffen wurde, und so den Weg für eine transparente und verifizierbare KI zu ebnen.

Schlüsselwörter Künstliche Intelligenz (KI) · Maschinelles Lernen (ML) · Interpretierbare KI · Verifizierbare KI · Konzept-Lernen

A. Holzinger (✉) · H. Müller
Medizinische Universität Graz, Auenbruggerplatz 2, 8036 Graz, Österreich
E-Mail: andreas.holzinger@human-centered.ai

A. Holzinger
xAI-Lab, Alberta Machine Intelligence Institute, Edmonton, Kanada

Combining Natural and Artificial Intelligence: An Experimental Test Environment for Explainable AI (xAI)

Abstract Artificial intelligence (AI) follows the concept of human intelligence, which unfortunately is not a clearly defined concept. The most common definition, as given in cognitive science as a mental ability, includes the ability to think abstract, logical and deductively and to solve given problems of the real world. A current topic in AI is to find out whether and to what extent algorithms are capable of learning such abstract “thinking” and reasoning similar to humans—or whether the learning outcome is based on purely statistical correlation. In this paper we present a freely available, universal and extensible experimental test environment. These “Kandinsky Patterns”, named after the Russian painter and art theorist Wassily Kandinsky (1866–1944), represent a kind of “Swiss knife” for studying the problems mentioned above. The area that deals with these problems is called “explainable AI” (xAI). Explainability/Interpretability aims to enable human experts to understand the underlying explanatory factors—causality—i.e. why an AI decision was made, thus paving the way for a transparent and verifiable AI.

Keywords Artificial Intelligence (AI) · Machine Learning (ML) · Explainable AI · Verifiable AI · Concept learning

1 Einführung und Motivation

Vor allem durch den großen Fortschritt des statistisch/probabilistischen maschinellen Lernens (ML) in den letzten Jahrzehnten wurden beeindruckende Erfolge erzielt. Die KI wurde damit nach dem langen KI-Winter (Hendler 2008) nicht nur wieder „salonfähig“, sondern sogar zum wirtschaftlichen Erfolg. Ob Spracherkennung (von Alexa bis Siri), Übersetzungssysteme (von Google Translate bis DeepL), Empfehlungssysteme (Recommender Systems), Klassifizierungsalgorithmen (Medizin), oder selbstfahrende Autos, erfolgreiche Beispiele gibt es in vielen wirtschaftlich hochrelevanten Anwendungsdomänen. Diese Erfolge sind aus Informatiksicht sehr eindrucksvoll und werden oft als Meilensteine zur Erreichung „menschlicher Intelligenz“ angesehen. Allerdings sind auch die derzeit erfolgreichsten Anwendungen nur in sehr eng begrenzten und sehr spezifischen Aufgaben erfolgreich (Dowe und Hernández-Orallo 2012), während Menschen einen Kontext verstehen und auch mit unvollständiger Information in komplexen, sich ändernden Problemstellungen umgehen können (Holzinger 2016b). Aber vielleicht das größte Problem der derzeit erfolgreichsten Lösungen ist deren mangelnde Nachvollziehbarkeit und Erklärbarkeit/Interpretierbarkeit (Holzinger 2018b).

Als praktisches Beispiel sei die hervorragende Arbeit in Nature aus der Stanford-Gruppe um Sebastian Thrun genannt (Esteva et al. 2017). In Europa wird diese Arbeit unter dem Titel „KI – so gut wie Hautärzte“ gerne als Paradebeispiel in der Medizin genannt. Es handelt sich dabei um eine Klassifizierung dermatologischer Bilder zur Identifizierung bösartiger Melanome. Der dazu verwendete „Deep Learning“ Ansatz in Form eines Inception v3 convolutional neural networks

(Krizhevsky et al. 2012), lieferte mit 92 % Performanz eindrucksvolle Ergebnisse. Allerdings werden solche Ansätze als „Black-Box“ bezeichnet und erlauben keine Nachvollziehbarkeit und damit keine Interpretierbarkeit und einer Möglichkeit zur Beantwortung der Frage *warum* diese 92 % bzw. 8 % Fehlklassifikationen entstanden sind. Insbesondere durch das „Recht auf Erklärung“ in der Europäischen Datenschutzgrundverordnung, steigt das Interesse in Forschung und Wirtschaft am Thema Erklärbarkeit/Interpretierbarkeit das unter dem Namen explainable AI (ex-AI oder kurz xAI) zusammengefasst wird (Holzinger 2018a). Die Nachvollziehbarkeit, Interpretation und Verstehbarkeit von Ergebnissen maschineller Lernalgorithmen und Möglichkeiten die *Qualität* von KI zu evaluieren rücken damit in das Zentrum des Interesses von Wissenschaft und Wirtschaft. Durch die Schwächen so-genannter „Black-Box“ Algorithmen sowie sozialer und ethischer Verantwortlichkeit ist man, zumindest in Domänen wie der Medizin, wieder stärker motiviert an einer Verbindung von menschlicher Intelligenz und maschineller Intelligenz („augmenting human intelligence with artificial intelligence“) zu arbeiten und die Stärken von KI und menschlicher Intelligenz zu verbinden (Holzinger 2016a; Holzinger et al. 2019b).

Im nachfolgenden zweiten Kapitel diskutieren wir zunächst einige theoretische Konzepte zum Begriff „Intelligenz“ und geben eine Beschreibung der Forschungslücke, warum die Kandinsky Patterns relevant sind. Im dritten Kapitel diskutieren wir einige derzeitige Möglichkeiten wie man Formen von „Intelligenz“ testen kann. Im vierten Kapitel schließlich stellen wir die Kandinsky Patterns vor.

2 Vom Begriff der Intelligenz zur erklärbaren KI

Ein grundlegendes Problem für die KI sind die oft vagen und sehr unterschiedlichen Definitionen des Begriffs Intelligenz. Dies ist bei künstlichen Systemen, die sich deutlich vom Menschen unterscheiden, besonders akut und auch seit längerer Zeit bekannt (Legg und Hutter 2007). Aus diesem Grund standen Intelligenztests für KI im Allgemeinen und ML im Besonderen bis dato nicht im Mittelpunkt umfangreicher Forschung in der internationalen KI/ML Gemeinschaft. Die Bewertung von Ansätzen und Algorithmen erfolgte in erster Linie anhand bestimmter Vergleichsmaßstäbe, so genannter „benchmarks“ (vgl. Nambiar et al. 2019; Nambiar 2018).

Der bekannteste Ansatz, der streng genommen keiner ist, ist jener von Alan Turing 1950 (Turing 1950): ein Algorithmus wird als intelligent (genug) für eine bestimmte Art von Aufgaben angesehen, wenn und nur wenn er alle möglichen Aufgaben seiner Art erledigen könnte. Der Mangel dieses Ansatzes besteht jedoch darin, dass er stark *aufgabenorientiert* ist und dass er ein hohes *a priori Wissen* über alle möglichen Aufgaben und die Möglichkeit, diese zu definieren, erfordert. Letzteres wiederum birgt das Problem der Granularität und Präzision von Definitionen. Ein indikatives Beispiel ist der „Intelligenztest“ für das autonome Fahren (Li et al. 2019). Ein weiteres, länger bekanntes Beispiel stellt der so-genannte CAPTCHA (vollständig automatisierter öffentlicher Turing-Test, um Computer und Menschen voneinander zu unterscheiden) dar, der einfach für Menschen, aber schwierig für Maschinen ist und daher für Sicherheitsanwendungen verwendet wird (Von Ahn et al. 2003). Solche

CAPTCHAs verwenden entweder Text oder Bilder unterschiedlicher Komplexität und stellen individuelle Unterschiede in der kognitiven Verarbeitung dar (Belk et al. 2013).

In der Kognitionswissenschaft hat die Prüfung der menschlichen Eignung – Intelligenz wird hier als eine Form der kognitiven Eignung gesehen – eine sehr lange Tradition. Grundsätzlich ergibt sich die Idee der psychologischen Messung aus den allgemeinen Entwicklungen in der Wissenschaft des 19. Jahrhunderts und insbesondere der Physik, die einen wesentlichen Schwerpunkt auf die genaue Messung von Variablen legt. Der Beginn menschlicher Intelligenztests erfolgte um 1900: Alfred Binet (1857–1911) begann mit der Entwicklung von Bewertungsfragen zur Identifizierung schulreifer Kinder. Bemerkenswert ist, dass Binet sich nicht nur auf Aspekte konzentrierte, die explizit in Schulen gelehrt wurden, sondern auch auf allgemeinere und abstraktere Fähigkeiten, einschließlich Aufmerksamkeitsspanne, Gedächtnis und Problemlösungsfähigkeiten. Binet und seine Kollegen fanden heraus, dass die Fähigkeit der Kinder, die Fragen zu beantworten und die Aufgaben zu lösen, nicht unbedingt eine Frage des körperlichen Alters war. Basierend auf dieser Beobachtung schlug Binet ein mentales Alter vor, das eigentlich die erste Intelligenzmessung darstellte. Das Niveau der Eignung wurde im Verhältnis zur durchschnittlichen Eignung der gesamten Bevölkerung gesehen. Im Laufe der Zeit entwickelten sich viele verschiedene Arten von Intelligenztests. Was wir in diesem Beitrag lediglich hervorheben möchten ist, dass diese sehr frühen Intelligenztests bereits den grundlegenden Unterschied zur aufgabenorientierten Auswertung der späteren KI sehr deutlich machen. **Die menschliche Intelligenz wurde nicht als die Fähigkeit angesehen, eine bestimmte Aufgabe zu lösen, wie beispielsweise eine reine Klassifikationsaufgabe, sondern als ein viel breiteres Konstrukt.** Darüber hinaus wurde die menschliche Intelligenz im Allgemeinen nicht isoliert gemessen, sondern immer in Bezug auf eine Grundgesamtheit. Am Beispiel der selbstfahrenden Autos wäre daher die Frage, ob ein Auto besser fahren kann als alle anderen Autos, oder ob und inwieweit das Auto besser ist als menschliche Fahrer.

Zum nachfolgenden Verständnis unserer Kandinsky Patterns erwähnen wir zwei grundlegende Arbeiten.

Die erste Arbeit ist jene von John Raven (1902–1970), der die sogenannten „Raven’s Progressive Matrices“ entwickelte. Dabei handelt es sich um eine **nonverbale** Multiple-Choice-Messung der Argumentationskomponente zweier Ausprägungen, namentlich (i) klares Denken und Komplexitätssinn und (ii) die Fähigkeit, Informationen zu speichern und zu reproduzieren (Raven 2000).

Die zweite Arbeit wurde von Mikhail Bongard (1924–1971) in Form sogenannter Bongard Probleme vorgestellt (Bongard 1967). Ein Bongard Problem besteht darin, zwei Mengen relativ einfacher Muster zu präsentieren, worin z. B. alle Muster aus Menge A einen gemeinsamen Faktor oder ein gemeinsames Attribut haben, das in allen Mustern aus Menge B fehlt. Das Problem besteht darin, den gemeinsamen Faktor zu finden oder zumindest überzeugend zu formulieren. Bongard Probleme wurden im bekannten Buch von Douglas Hofstadter (geb. 1945) beschrieben (Hofstadter 1979), hatten aber interessanterweise kaum Einfluss auf die KI. Die Theorie die diesen Ansätzen zugrunde liegt ist das „Lernen von Konzepten“ (Hunt 1962; Lake et al. 2015; Mao et al. 2015).

Im Bereich der „erklärbaren KI“ (üblicher ist die angelsächsische Bezeichnung explainable AI) werden Methoden entwickelt, um insbesondere Deep-Learning Modelle interpretierbar zu machen. Dies kann z.B. mit Hilfe einfacher Sensitivitätsanalyse (Hampel et al. 2011) erfolgen, um die Vorhersage in Bezug auf die Eingangsgrößen zu verstehen. Eine sehr bekannte und erfolgreiche Methode ist Layer-Wise-Relevance Propagation (LRP), wobei mit einer Heatmap sehr gut dargestellt werden kann welche Eingangsparameter am meisten zum Ergebnis beitragen (Bach et al. 2015; Lapuschkin et al. 2016)¹. Die Heatmap-Visualisierung, siehe z. B. (Sturm et al. 2015) zeigt z.B. an, welche Pixel verändert werden müssen, damit das Bild (aus Sicht der KI-Systeme!) etwa wie die vorhergesagte Klasse aussieht.

Der Schlüssel zu einer effektiven Mensch-KI-Interaktion und daher auch der Erfolg zukünftiger Human-AI Interfaces liegt in einer effizienten und übereinstimmenden Abbildung („mapping“) der Erklärbarkeit („Explainability“ durch „künstliche Intelligenz“) mit der Kausalität im Sinne von (Pearl 2009) (durch „menschliche Intelligenz“).

Bei dieser „Landkartenmetapher“ geht es darum, Verbindungen und Beziehungen zwischen Bereichen herzustellen – und *nicht* (!) darum, eine neue Karte zu zeichnen, vielmehr geht es darum, die gleichen (oder zumindest ähnliche) Gebiete in zwei völlig unterschiedlichen „Karten“ zu identifizieren – deswegen ist „mapping“ ein sehr guter Begriff. Ein effektives mapping ist notwendig, aber nicht hinreichend für das Verständnis einer Erklärung. Ob eine Erklärung verstanden worden ist hängt von weiteren Faktoren (wie z.B. Vorwissen) ab. Wenn die Erklärung gut war hat man etwas verstanden. Um langfristige wirtschaftliche Erfolge in verschiedenen Anwendungsdomänen zu gewährleisten werden in Zukunft neue Human-AI Interfaces notwendig werden, die ein beständiges Feedback erlauben ob etwas verstanden wurde oder nicht. In einer Mensch-zu-Mensch Interaktion wird dieses Feedback sehr stark über Gesichtsausdrücke (emotion) sichergestellt, daher wird in Zukunft das Thema „emotion“ (Picard et al. 2001; Stickel et al. 2009), und emotional interfaces einen wichtigen Stellenwert für explainable AI bekommen.

Weiterhin ist es wichtig, zwischen Explainability und Causability zu unterscheiden. Unter Explainability (wie schon oben erwähnt) verstehen wir die Eigenschaft eines KI-Systems, „maschinelle Erklärung“ (z. B. durch „heatmapping“ (Sturm et al. 2015; Bach et al. 2015)) zu generieren. Causability, in Anlehnung an das Wort Usability, ist die *Qualität* mit der eine solche Abbildung erfolgt, also ein mapping wie auf einer Landkarte zwischen KI und dem Menschen; d.h. zwischen der „maschinellen Erklärung“ und der Eigenschaft von Menschen Erklärungen zu verstehen (Holzinger et al. 2019a). Um die Qualität solcher Erklärungen zu bewerten haben wir eine System Causability Scale (SCS) entwickelt (Holzinger et al. 2020).

3 Über das Testen von Intelligenz

Innerhalb der Machine Learning Community wird intensiv diskutiert, ob z.B. neuronale Netze abstraktes Denken lernen können oder ob sie sich nur auf reine Kor-

¹ Sebastian Bach heißt seit 2016 Sebastian Lapuschkin, es handelt sich also um denselben Autor.

relationen verlassen. In einem kürzlich erschienenen Beitrag schlagen die Autoren (Santoro et al. 2018) einen Datensatz und eine challenge vor, um das abstrakte Denken zu untersuchen, das von einem bekannten menschlichen IQ-Test inspiriert ist: dem Raven-Test, genauer gesagt den Ravens Progressive Matrices (RPM) und den Mill Hill Vocabulary Scales, die 1936 für die Grundlagenforschung sowohl zu den genetischen als auch zu den Umweltfaktoren der „Intelligenz“ entwickelt wurden, wie bereits in Abschn. 3 erwähnt (Raven 2000). Die Prämisse hinter RPMs ist einfach: Man muss über die Beziehungen zwischen wahrnehmbaren visuellen Merkmalen (wie z. B. Formpositionen oder Linienfarben) nachdenken, um ein Bild zu wählen, das die Matrix vervollständigt. So nimmt beispielsweise die Größe der Quadrate entlang der Zeilen zu, und das richtige Bild ist das, das sich an dieses Größenverhältnis hält.

RPMs sind eine starke Diagnostik der abstrakten verbalen, räumlichen und mathematischen Denkfähigkeit. Um der Herausforderung erfolgreich zu begegnen, müssen Modelle mit verschiedenen Verallgemeinerungsregimen umgehen, bei denen sich die Trainings- und Testdaten klar voneinander unterscheiden. Das erstaunlich fortschreitende Feld der KI- und ML-Technologien fügt dem Diskurs der Intelligenztests eine weitere Dimension hinzu, nämlich die Bewertung der künstlichen Intelligenz im Gegensatz zur menschlichen Intelligenz. Die menschliche Intelligenz neigt dazu, sich auf der Grundlage verschiedener kognitiver und neuronaler Prozesse an die Umgebung anzupassen. Der Bereich der KI wiederum konzentriert sich sehr stark auf die Entwicklung von Algorithmen, die das menschliche Verhalten nachahmen können (schwache oder enge KI).

Dies gilt insbesondere für angewandte Genres wie autonomes Fahren von Autos, Robotik oder Spielen. Dies führt auch zu deutlichen Unterschieden in dem, was wir für intelligent halten. Der Mensch hat ein Bewusstsein, er kann improvisieren, und die menschliche Physiologie zeigt Plastizität, die zu echtem Lernen führt, indem der Mensch dieses „Bewusstsein“ selbst „verändern“ kann. **Obwohl Menschen dazu neigen, mehr Fehler zu machen, ist die menschliche Intelligenz als solche in der Regel zuverlässiger und robuster gegen katastrophale Fehler, während die KI anfällig für bereits kleine Störungen ist, wie z. B. für Softwarefehler, Hardware- und Energieausfälle.**

Die menschliche Intelligenz entwickelt sich auf der Grundlage unendlicher Interaktionen mit einer unendlichen Umgebung, während die KI auf die kleine Welt einer bestimmten Aufgabe beschränkt ist.

Wir wollen diese Idee durch die Herausforderung der Identifikation und Interpretation/Erklärung visueller Muster veranschaulichen. Im Wesentlichen bezieht sich dies auf die menschliche Fähigkeit, den Sinn der Welt zu verstehen (z. B. durch die Identifizierung der Natur einer Reihe von visuellen Mustern, die fortgesetzt werden müssen). Sensemaking ist eine aktive Verarbeitung von Empfindungen, um ein Verständnis der Außenwelt zu erlangen, und beinhaltet den Erwerb von Informationen, das Lernen über neue Bereiche, die Lösung von Problemen, den Erwerb von Situationsbewusstsein und die Teilnahme am sozialen Wissensaustausch (Pirulli und Russell 2011). Die Fähigkeit kann auf konkrete Bereiche wie verschiedene Human-Computer Interaktionen, aber auch auf abstrakte Bereiche wie die Mustererkennung angewendet werden. Dieses Thema stand speziell im Fokus der medizinischen For-

schung. Kundel und Nodine (Kundel und Nodine 1983) untersuchten beispielsweise Blickpfade in medizinischen Bildern (ein Sonogramm, ein Tomogramm und zwei Standard-Röntgenbilder) – ähnlich wie in modernen Arbeiten (Pohn et al. 2019). Die Versuchspersonen wurden gebeten, jedes der Bilder in einem Satz zusammenzufassen. Die Ergebnisse dieser Studie zeigten, dass die korrekte Interpretation der Bilder im Zusammenhang mit dem Besuch der relevanten Bereiche der Bilder stand und nicht mit dem Besuch visuell dominanter Bereiche der Bilder. Die Autoren fanden auch einen starken Zusammenhang von Erklärungen zu Erfahrungen mit Bildern. Ein Grundprinzip in der Wahrnehmung und Interpretation visueller Muster ist das ursprünglich von Helmholtz formulierte Wahrscheinlichkeitsprinzip, das besagt, dass die bevorzugte Wahrnehmungsorganisation eines abstrakten visuellen Musters auf der Wahrscheinlichkeit bestimmter Objekte basiert (Leeuwenberg und Boselie 1988).

Eine bis zu einem gewissen Grad konkurrierende Erklärung ist das von der Gestaltpsychologie vorgeschlagene Mindestprinzip, das behauptet, dass der Mensch ein visuelles Muster nach möglichst einfacher Interpretation wahrnimmt. Die Rolle der Erfahrung spiegelt sich auch in Studien im Kontext der Wahrnehmung von abstrakter versus repräsentativer bildender Kunst wider; (Uusitalo et al. 2009) zeigte deutliche Unterschiede bei Kunstexperten und Laien in der Wahrnehmung und ihren Präferenzen der bildenden Kunst. Psychologische Untersuchungen konnten zeigen, dass die Art der Wahrnehmung und Interpretation visueller Muster daher eine Funktion der Erwartungen ist (Yanagisawa 2019). Dies führt einerseits oft zu Fehlinterpretationen oder vorzeitigen Interpretationen, andererseits erhöht es die Erklärbarkeit von Interpretationen, da die visuelle Wahrnehmung durch bestehende Konzeptualisierungen bestimmt wird.

4 Kandinsky Patterns

Zur Entwicklung unserer Kandinsky Patterns wurden wir inspiriert durch unsere Erfahrung in der Zusammenarbeit mit Pathologinnen und Pathologen. Diese beschreiben z. B. histopathologische Bilder indem sie geometrische Objekte identifizieren, sie sprechen von Architekturen und identifizieren Regelmäßigkeiten und Anomalien dieser geometrischen Strukturen. Im ersten Schritt beschreiben sie was sie sehen, in einem zweiten Schritt interpretieren sie ihre Beobachtungen. Wird KI/ML auf solche digitalen Bilder angewandt, kommt man sehr rasch an ein großes Problem: Fehlende Grundwahrheit („Ground truth“). Dieses Problem war ein zentraler Motivator für die Entwicklung der Kandinsky Patterns.

Kandinsky Patterns (Müller und Holzinger 2019)² sind mathematisch beschreibbare, einfache, in sich geschlossene und damit mathematisch streng kontrollierbare Testdatensätze (Bilder) für die Entwicklung, Validierung und das Training der Erklärbarkeit in der KI. Kandinsky Patterns sind allerdings gleichzeitig (!) leicht von Menschen zu unterscheiden und daher können kontrollierte Muster sowohl von Menschen als auch von Algorithmen beschrieben/verarbeitet werden. Dies ist sehr

² <https://human-centered.ai/project/kandinsky-patterns/>.

wichtig, um Erklärungsprozesse von Algorithmen mit denen von Menschen zu vergleichen und zu verstehen. Damit gewinnen wir grundlegende Erkenntnisse für das Gebiet der explainable AI. Das wichtigste dabei ist, dass wir die „Ground Truth“ („Goldstandard“ fehlt in der real-world leider sehr oft) zu jeder Zeit generieren, beim Testen verbergen, aber stets vergleichen können.

Der Begriff „Ground Truth“ (eigentlich „Bodenwahrheit“) entstammt ursprünglich aus der Kartographie und Fernerkundung (Geographical Information Systems, GIS), wo nur durch das Vorhandensein von sogenannten „Ground-Truth Daten“, es möglich wird ein Ergebnis auf Korrektheit zu prüfen (Pickles 1995).

Die Ergebnisse einer ersten, kürzlich durchgeführten Studie über das Erklärungsverhalten von Menschen (Holzinger et al. 2019c) zeigten, dass die Mehrheit der Erklärungen auf der Grundlage der *Eigenschaften* der einzelnen Elemente in einem Bild (d.h. Form, Farbe, Größe) und des Aussehens der einzelnen Objekte (Anzahl) gemacht wurde. Vergleiche von Elementen (z. B. mehr, weniger, größer, kleiner, etc.) waren deutlich unwahrscheinlicher und die Lage von Objekten spielte interessanterweise fast keine Rolle bei der Erklärung der Bilder.

In einer natürlichsprachlichen Aussage über eine Kandinsky-Figur verwenden Menschen eine Reihe von **Grundkonzepten**, die durch logische Operatoren kombiniert werden. Die folgenden (unvollständigen) Beispiele veranschaulichen einige Konzepte mit zunehmender Komplexität (siehe Abb. 1).

- Grundlegende Konzepte, die durch die Definition einer Kandinsky-Figur gegeben sind: eine Reihe von Objekten, beschrieben durch Form, Farbe, Größe und Position, siehe Abb. 1a–d für Farbe und Abb. 1e–h für Formen.
- Vorhandensein, Zahlen, Mengenverhältnisse (Anzahl, Menge oder Mengenverhältnisse von Objekten), z. B. „eine Kandinsky-Figur enthält 4 rote Dreiecke und mehr gelbe Objekte als Kreise“, siehe Abb. 1i–l.
- Raumkonzepte, die die Anordnung von Objekten beschreiben, entweder absolut (oben, oben, unten, links, rechts, ...) oder relativ (oben, oben, oben, berührend, ...), z. B. „in einer Kandinsky-Figur befinden sich rote Objekte auf der linken Seite, blaue Objekte auf der rechten Seite und gelbe Objekte auf der rechten Seite unter blauen Feldern“, siehe Abb. 1m–p.
- Gestaltkonzepte (siehe unten) z. B. Schließung, Symmetrie, Kontinuität, Nähe, Ähnlichkeit, z. B. „In einer Kandinsky-Figur werden Objekte kreisförmig gruppiert“, siehe Abb. 1q–t.
- Domänenkonzepte, z. B. „eine Gruppe von Objekten wird als „Blume“ wahrgenommen“, siehe Abb. 1u–x.

Diese Grundkonzepte können verwendet werden, um Gruppen von Objekten auszuwählen, z. B. „alle roten Kreise in der linken oberen Ecke“, und um einzelne Objekte und Gruppen in einer Anweisung mit logischem Operator weiter zu kombinieren, z. B. „wenn es einen roten Kreis in der linken oberen Ecke gibt, gibt es kein blaues Objekt“, oder mit komplexen domänenspezifischen Regeln, z. B. „wenn die Größe eines roten Kreises kleiner ist als die Größe eines gelben Kreises, sind rote Kreise kreisförmig um gelbe Kreise angeordnet“.

Siehe nachfolgendes Bild.

Abb. 1 Kandinsky-Muster, das Konzepte wie Farbe (a–d), Form (e–h), numerische Beziehungen (i–l), räumliche Beziehungen (m–p), Gestalt Konzepte (q–t) und Domänenkonzepte (u–x) zeigt

In ihren Experimenten (Hubel und Wiesel 1962) entdeckten sie unter anderem, dass das menschliche visuelle System ein Bild von sehr einfachen Reizen zu komplexeren Darstellungen aufbaut. Dies inspirierte die neuronale Netzwerkgemeinschaft, ihre so genannten „Deep Learning“-Modelle als kaskadierendes Modell von Zelltypen zu sehen, das immer ähnlichen einfachen Regeln folgt: Zuerst werden Linien

gelernt, dann Formen, dann Objekte gebildet, die schließlich zu Konzeptrepräsentationen führen (Schmidhuber 2015).

Durch die Verwendung von „Backpropagation“ (Lecun et al. 1989) ist ein solches Modell in der Lage, komplizierte Strukturen in großen Datensätzen zu entdecken, um anzuzeigen, wie die internen Parameter angepasst werden sollten, die verwendet werden, um die Darstellung in jeder Schicht aus der Darstellung in der vorherigen Schicht zu berechnen. Die Darstellung von Konzepten bezieht sich auf die menschliche Fähigkeit, Kategorien für Objekte zu lernen und neue Instanzen dieser Kategorien zu erkennen. Im maschinellen Lernen wird Konzeptlernen definiert als die Ableitung einer booleschen Funktion aus Trainingsbeispielen ihrer Ein- und Ausgänge, d.h. es trainiert einen Algorithmus zur Unterscheidung zwischen Beispielen und Nicht-Beispielen – wir nennen diese „counterfactuals“ („Gegenfakten“, „was wäre, wenn ...“). Konzeptlernen ist seit langem ein relevantes Forschungsgebiet im Bereich des maschinellen Lernens und hat seinen Ursprung in der Kognitionswissenschaft, definiert als die Suche nach Attributen, mit denen sich Exemplare von Nichtmustern verschiedener Kategorien unterscheiden lassen (Bruner 1956).

Die Fähigkeit, in Abstraktionen zu denken, ist eines der mächtigsten Werkzeuge, die Menschen besitzen. Technisch ordnen Menschen ihre Erfahrungen in kohärente Kategorien ein, indem sie eine gegebene Situation als Mitglied dieser Sammlung von Situationen definieren, für die Antworten x, y, etc. höchstwahrscheinlich angemessen sind. Diese Klassifizierung ist kein passiver Prozess und zu verstehen, wie Menschen Abstraktionen lernen, ist nicht nur für das Verständnis des menschlichen Denkens, sondern auch für den Bau von Maschinen der künstlichen Intelligenz unerlässlich (Hunt 1962).

Insbesondere bei lernenden Klassifikationsmodellen für die Segmentierung geht es darum, zwischen „guten“ und „schlechten“ Segmentierungen zu klassifizieren und die Gestalt-cues als Merkmale (die Prioren) zum Trainieren des Lernmodells zu verwenden. Vom Menschen manuell segmentierte Bilder werden als Beispiele für „gute“ Segmentierungen verwendet (Grundwahrheit), und „schlechte“ Segmentierungen werden durch zufällige Zuordnung einer menschlichen Segmentierung zu einem anderen Bild konstruiert (Ren und Malik 2003).

Gestalt-Prinzipien (Koffka 1935) können als Regeln angesehen werden, d.h. sie diskriminieren konkurrierende Segmentierungen nur, wenn alles andere gleich ist, daher sprechen wir allgemeiner als Gestaltgesetze und eine bestimmte Gruppe von Gestaltgesetzen sind die Gestaltgesetze der Gruppierung, genannt Prägnanz (Wertheimer 1938), die das Gesetz der Nähe beinhalten: Objekte, die nahe beieinander liegen, scheinen Gruppen zu bilden, auch wenn sie völlig unterschiedlich sind, das Gesetz der Ähnlichkeit: ähnliche Objekte werden gruppiert; oder das Gesetz des Verschlusses: Objekte können als solche wahrgenommen werden, auch wenn sie unvollständig oder durch andere Objekte verborgen sind. Wie eingangs erwähnt haben leider die derzeit leistungsfähigsten maschinellen Lernmethoden eine Reihe von Nachteilen, von denen eine besonders relevant ist, eben jene dass „Black-Box“ Ansätze aufgrund ihrer Komplexität schwer zu interpretieren. Bildklassifizierer arbeiten nicht mit High-Level-Konzepten, sondern mit Low-Level-Features (z. B. Linien, Kreise, etc.) und mit Domänenkonzepten, und gerade dies macht ihr Innen-

leben schwer zu interpretieren und zu verstehen. Das „Warum“ ist jedoch oft viel nützlicher als das reine Klassifizierungsergebnis.

5 Zusammenfassung

Durch den Vergleich sowohl der Stärken der maschinellen Intelligenz als auch der menschlichen Intelligenz ist es möglich, Probleme zu lösen, bei denen uns derzeit geeignete Methoden fehlen. Eine große allgemeine Frage ist: „Wie können wir eine Aufgabe erfüllen, indem wir das bei der Lösung früherer Aufgaben gewonnene Wissen nutzen?“ Um solche Fragen zu beantworten, ist es notwendig, Einblick in das menschliche Erklärungsverhalten zu gewinnen, aber nicht mit dem Ziel, menschliches Verhalten nachzuahmen, sondern menschliche Lernmethoden mit maschinellen Lernmethoden zu vergleichen. Wir hoffen, dass unsere Kandinsky Patterns die internationale Machine Learning Community herausfordern und freuen uns auf viele Kommentare und Ergebnisse. Um den Schritt zu einer menschenähnlicheren und wahrscheinlich gründlicheren Bewertung von KI/ML zu machen, schlagen wir vor, die Prinzipien menschlicher Intelligenztests anzuwenden, wie in diesem Beitrag beschrieben. Zur Beurteilung von KI/ML kann eine Reihe von Kandinsky-Mustern verwendet werden, von denen jedes für sich genommen komplex ist. Eine „echte“ intelligente Leistung wäre die Identifizierung der Konzepte – und damit der Bedeutung! – von Sequenzen mehrerer Kandinsky-Muster. Jedenfalls sind noch viel mehr experimentelle und theoretische Arbeiten erforderlich.

Danksagung Die Autoren bedanken sich sehr herzlich für die konstruktiven Kommentare der anonymen Gutachter. Teile dieser Arbeit wurden aus dem Projekt FWF P-32554 „Ein Referenzmodell für explainable AI“ gefördert.

Funding Open access funding provided by Austrian Science Fund (FWF).

Open Access Dieser Artikel wird unter der Creative Commons Namensnennung 4.0 International Lizenz veröffentlicht, welche die Nutzung, Vervielfältigung, Bearbeitung, Verbreitung und Wiedergabe in jeglichem Medium und Format erlaubt, sofern Sie den/die ursprünglichen Autor(en) und die Quelle ordnungsgemäß nennen, einen Link zur Creative Commons Lizenz beifügen und angeben, ob Änderungen vorgenommen wurden.

Die in diesem Artikel enthaltenen Bilder und sonstiges Drittmaterial unterliegen ebenfalls der genannten Creative Commons Lizenz, sofern sich aus der Abbildungslegende nichts anderes ergibt. Sofern das betreffende Material nicht unter der genannten Creative Commons Lizenz steht und die betreffende Handlung nicht nach gesetzlichen Vorschriften erlaubt ist, ist für die oben aufgeführten Weiterverwendungen des Materials die Einwilligung des jeweiligen Rechteinhabers einzuholen.

Weitere Details zur Lizenz entnehmen Sie bitte der Lizenzinformation auf <http://creativecommons.org/licenses/by/4.0/deed.de>.

Literatur

Von Ahn L, Blum M, Hopper NJ, Langford J (2003) CAPTCHA: using hard AI problems for security. In: Bilham E (Hrsg) Advances in Cryptology-Eurocrypt 2003. Springer, Berlin, S 294–311 https://doi.org/10.1007/3-540-39200-9_18

- Bach S, Binder A, Montavon G, Klauschen F, Müller K-R, Samek W (2015) On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PLoS ONE* 10(7):e130140. <https://doi.org/10.1371/journal.pone.0130140>
- Belk M, Germanakos P, Fidas C, Holzinger A, Samaras G (2013) Towards the personalization of CAPTCHA mechanisms based on individual differences in cognitive processing. In: *Human factors in computing and Informatics. Lecture notes in computer science*, LNCS 7946. Springer, Berlin Heidelberg, S 409–426 https://doi.org/10.1007/978-3-642-39062-3_26
- Bongard MM (1967) The problem of recognition. Nauka, Moscow (in Russian)
- Bruner JS (1956) On attributes and concepts. In: Bruner JS, Goodnow JJ, Austin GA (Hrsg) *A study of thinking*. Wiley, New York, S 25–49
- Dowe DL, Hernández-Orallo J (2012) IQ tests are not for machines, yet. *Intelligence* 40(2):77–81. <https://doi.org/10.1016/j.intell.2011.12.001>
- Esteva A, Kuprel B, Novoa RA, Ko J, Swetter SM, Blau HM, Thrun S (2017) Dermatologist-level classification of skin cancer with deep neural networks. *Nature* 542(7639):115–118. <https://doi.org/10.1038/nature21056>
- Hampel FR, Ronchetti EM, Rousseeuw PJ, Stahel WA (2011) *Robust statistics: the approach based on influence functions*. Wiley, New York
- Hendler J (2008) Avoiding another AI winter. *IEEE Intell Syst* 23(2):2–4. <https://doi.org/10.1109/MIS.2008.20>
- Hofstadter DR (1979) *Goedel, Escher, Bach: an eternal golden braid*. Basic Books, New York
- Holzinger A (2016a) Interactive machine learning (iML). *Inform Spektrum* 39(1):64–68. <https://doi.org/10.1007/s00287-015-0941-6>
- Holzinger A (2016b) Interactive machine learning for health Informatics: when do we need the human-in-the-loop? *Brain Inform* 3(2):119–131. <https://doi.org/10.1007/s40708-016-0042-6>
- Holzinger A (2018a) Explainable AI (ex-AI). *Inform Spektrum* 41(2):138–143. <https://doi.org/10.1007/s00287-018-1102-5>
- Holzinger A (2018b) Interpretierbare KI: Neue Methoden zeigen Entscheidungswege künstlicher Intelligenz auf. *c't* 22:136–141
- Holzinger A, Carrington A, Müller H (2020) Measuring the quality of explanations: the system causability scale (SCS). *Künstl Intell*. <https://doi.org/10.1007/s13218-020-00636-z>
- Holzinger A, Kickmeier-Rust M, Müller H (2019c) KANDINSKY patterns as IQ-test for machine learning. In: *International cross-domain conference for machine learning and knowledge extraction. Lecture notes in computer science LNCS 11713*. Springer, Cham, S 1–14 https://doi.org/10.1007/978-3-030-29726-8_1
- Holzinger A, Langs G, Denk H, Zatloukal K, Mueller H (2019a) Causability and explainability of artificial intelligence in medicine. *Wiley Interdiscip Rev Data Min Knowl Discov*. <https://doi.org/10.1002/widm.1312>
- Holzinger A, Plass M, Kickmeier-Rust M, Holzinger K, Crişan GC, Pintea C-M, Palade V (2019b) Interactive machine learning: experimental evidence for the human in the algorithmic loop. *Appl Intell* 49(7):2401–2414. <https://doi.org/10.1007/s10489-018-1361-5>
- Hubel DH, Wiesel TN (1962) Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *J Physiol* 160(1):106–154. <https://doi.org/10.1113/jphysiol.1962.sp006837>
- Hunt EB (1962) *Concept learning: an information processing problem*. Wiley, Hoboken <https://doi.org/10.1037/13135-001>
- Koffka K (1935) *Principles of gestalt psychology*. Harcourt, New York
- Krizhevsky A, Sutskever I, Hinton GE (2012) Imagenet classification with deep convolutional neural networks. In: Pereira F, Burges CJC, Bottou L, Weinberger KQ (Hrsg) *Advances in neural information processing systems (NIPS 2012)*. NIPS, Lake Tahoe, S 1097–1105
- Kundel HL, Nodine CF (1983) A visual concept shapes image perception. *Radiology* 146(2):363–368
- Lake BM, Salakhutdinov R, Tenenbaum JB (2015) Human-level concept learning through probabilistic program induction. *Science* 350(6266):1332–1338. <https://doi.org/10.1126/science.aab3050>
- Lapuschkin S, Binder A, Montavon G, Müller K-R, Samek W (2016) The LRP toolbox for artificial neural networks. *J Mach Learn Res* 17(1):3938–3942
- Lecun Y, Boser B, Denker JS, Henderson D, Howard RE, Hubbard W, Jackel LD (1989) Backpropagation applied to handwritten zip code recognition. *Neural Comput* 1(4):541–551. <https://doi.org/10.1162/neco.1989.1.4.541>
- Leeuwenberg EL, Boselie F (1988) Against the likelihood principle in visual form perception. *Psychol Rev* 95(4):485–491. <https://doi.org/10.1037/0033-295X.95.4.485>

- Legg S, Hutter M (2007) Universal intelligence: a definition of machine intelligence. *Minds Mach* 17(4):391–444. <https://doi.org/10.1007/s11023-007-9079-x>
- Li L, Wang X, Wang K, Lin Y, Xin J, Chen L, Xu L, Tian B, Ai Y, Wang J (2019) Parallel testing of vehicle intelligence via virtual-real interaction. *Sci Robot* 4(eaaw4106):1–3
- Mao J, Wei X, Yang Y, Wang J, Huang Z, Yuille AL (2015) Learning like a child: fast novel visual concept learning from sentence descriptions of images. In: *Proceedings of the IEEE international conference on computer vision ICCV 2015*.
- Müller H, Holzinger A (2019) Kandinsky patterns. *arXiv:1906.00657*
- Nambiar R (2018) Towards an industry standard for benchmarking artificial intelligence systems. In: *IEEE 34th international conference on data engineering ICDE 2018*. <https://doi.org/10.1109/ICDE.2018.00212>
- Nambiar R, Ghandeharizadeh S, Little G, Boden C, Dholakia A (2019) Industry panel on defining industry standards for benchmarking artificial intelligence. *Springer, Cham*, S 1–6 https://doi.org/10.1007/978-3-030-11404-6_1
- Pearl J (2009) *Causality: models, reasoning, and inference*, 2. Aufl. Cambridge University Press, Cambridge
- Picard RW, Vyzas E, Healey J (2001) Toward machine emotional intelligence: analysis of affective physiological state. *IEEE Trans Pattern Anal Machine Intell* 23(10):1175–1191
- Pickles J (1995) *Ground truth: the social implications of geographic information systems*. Guilford, New York, London
- Pirolli P, Russell DM (2011) Introduction to this special issue on sensemaking. *Hum Comput Interact* 26(1/2):1–8. <https://doi.org/10.1080/07370024.2011.556557>
- Pohn B, Mayer M-C, Reihs R, Holzinger A, Zatloukal K, Müller H (2019) Visualization of histopathological decision making using a roadbook metaphor. In: *23rd international conference information visualisation (IV)* <https://doi.org/10.1109/IV.2019.00073>
- Raven J (2000) The Raven's progressive matrices: change and stability over culture and time. *Cogn Psychol* 41(1):1–48. <https://doi.org/10.1006/cogp.1999.0735>
- Ren X, Malik J (2003) Learning a classification model for segmentation. In: *Ninth IEEE international conference on computer vision (ICCV)* <https://doi.org/10.1109/ICCV.2003.1238308>
- Santoro A, Hill F, Barrett D, Morcos A, Lillicrap T (2018) Measuring abstract reasoning in neural networks. In: *35th international conference on machine learning*. PMLR, Stockholm, S 4477–4486
- Schmidhuber J (2015) Deep learning in neural networks: an overview. *Neural Netw* 61(1):85–117. <https://doi.org/10.1016/j.neunet.2014.09.003>
- Stickel C, Ebner M, Steinbach-Nordmann S, Searle G, Holzinger A (2009) Emotion detection: application of the valence arousal space for rapid biological usability testing to enhance universal access. In: Stephanidis C (Hrsg) *Universal access in human-computer interaction. Addressing diversity*. Lecture Notes in Computer Science, LNCS 5614. Springer, Berlin, Heidelberg, S 615–624 https://doi.org/10.1007/978-3-642-02707-9_70
- Sturm W, Schaefer T, Schreck T, Holzinger A, Ullrich T (2015) Extending the scaffold Hunter visualization toolkit with interactive heatmaps. In: Borgo R, Cagatay T (Hrsg) *EG UK computer graphics & visual computing CGVC 2015*, S 77–84 <https://doi.org/10.2312/cgvc.20151247>
- Turing AM (1950) Computing machinery and intelligence. *Mind* 59(236):433–460. <https://doi.org/10.1093/mind/LIX.236.433>
- Uusitalo L, Simola J, Kuisma J (2009) Perception of abstract and representative visual art. In: *Proceedings of AIMAC, 10th conference of the international association of arts and cultural management*
- Wertheimer M (1938) Laws of organization in perceptual forms. In: Ellis WD (Hrsg) *A source book of Gestalt psychology (with an introduction by K. Koffka)*. Paul Kegan, London, S 71–88 <https://doi.org/10.1037/11496-005>
- Yanagisawa H (2019) How does expectation affect sensory experience? A theory of relativity in perception. In: *5th international symposium on affective science and engineering ISASE 2019*. Japan Society of Kansei Engineering, Kogakuin, S 1–4 <https://doi.org/10.5057/isase.2019-C000014>