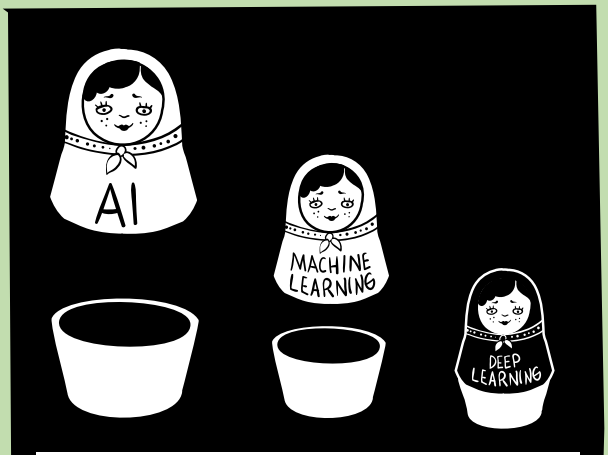
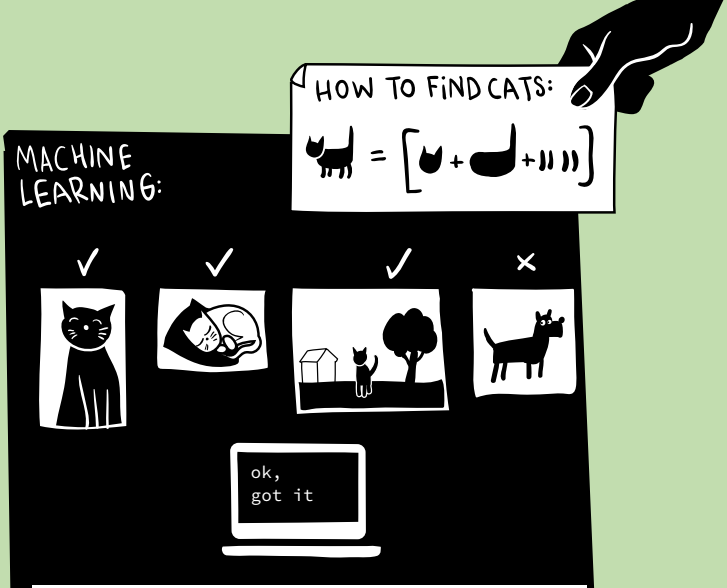
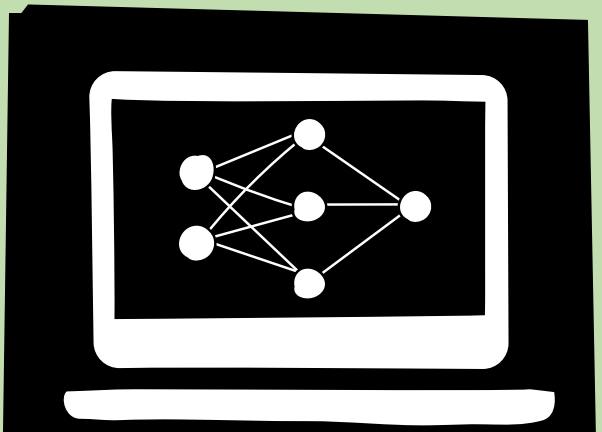


What do these buzzwords even mean? Machine Learning; Deep Learning; Neural Networks (NN)...?

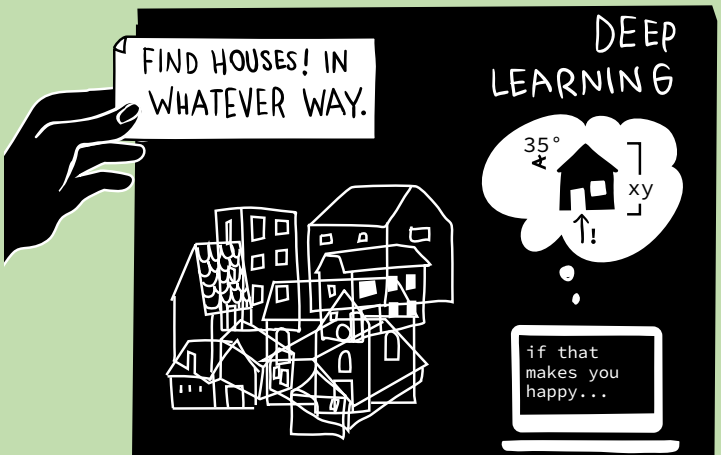


Machine Learning (ML) helps recognizing patterns based on existing data and algorithms: an important branch of AI.

(Note: The AI bookshelf includes other topics, as well.)



Before Deep Learning (DL), we had to know a lot about the data we were programming. Otherwise, ML would break down.



DL uses artificial Neural Networks (NN) that learn patterns directly from the data they are fed.

Like our brain, NNs compare new information with objects they know. However, they have not yet been able to think of new objects.

www.weneedtotalk.ai

Stephan Dreyer

Auch im Medienbereich haben Algorithmen, Formen maschinellen Lernens und Automatisierung längst Einzug gehalten. Der Jugendmedienschutz, der bei praktisch allen Medienangeboten immer mitschwingt, ist dabei eine rechtlich relevante Querschnittsmaterie, die bislang noch wenig in Erscheinung getreten ist, wenn es um den Einsatz künstlicher Intelligenz (KI) geht. Der Beitrag gibt einen Überblick, wo Automatisierung im Jugendschutz bereits Fuß gefasst hat oder Thema von Zukunftsplänen ist, und zeigt auf, was die künstliche Intelligenz in welchen Einsatzgebieten schon kann – und wo die Grenzen liegen.

Künstlich, intelligent, hilfreich?

Einsatzorte von künstlicher Intelligenz im Jugendmedienschutz

Über *lernende* Maschinen und *intelligente* Software

Im Zentrum der künstlichen Intelligenz stehen derzeit vor allem Technologien maschinellen Lernens, d. h. mathematische Verfahren, die Muster und Korrelationen in Daten erkennen und so auf Grundlage von Trainingsdaten statistische Modelle berechnen, welche nach dem Training auch auf unbekannte Daten angewandt werden. Das System erkennt dann etwa bestimmte Gegenstände, Geräusche, Personen oder Aktionen, versteht gesprochene Wörter, Sätze und Themen oder kann statistische Auffälligkeiten im Rahmen der Nutzung, Bewertung oder des Teilens einzelner Inhalte identifizieren. Auf dieser Grundlage kann der Algorithmus Softwareentscheidungen treffen, z. B. ein vom Nutzer geladenes Video mit Inhaltsbegriffen verschlagworten und die Untertitel automatisiert erstellen, Fotos bekannten Personen, Profilen oder Objektkategorien zuordnen, zudem Aussagen über die Wahrscheinlichkeit eines bestimmten Nutzungsverlaufs machen. Wirklich *intelligent* ist die Software dabei nicht: Sie versteht nicht den Sinn der Worte oder den sozialen Kontext von identifizierten Objekten, sie kennt keine Konzepte wie Recht oder Ethik oder Entwicklungsbeeinträchtigung. Sie errechnet auf Grundlage eines statistischen Modells mathematische Annahmen.

Klassisches Einsatzgebiet: Wiedererkennung von bekannten Inhalten

Ein bereits seit Langem bestehendes Einsatzgebiet für automatisierte Verfahren im Kinder- und Jugendschutz[1] ist die Identifizierung von Darstellungen sexuellen Kindesmissbrauchs („child sexual abuse" [CSA]). Während die erste Generation derartiger Techniken (z. B. PhotoDNA) vor allem einmal identifiziertes Material auf Basis von dateispezifischen Hashwerten wiedererkennen konnte – und damit genau genommen nicht auf Basis künstlicher Intelligenz arbeitete –, zielen neuere Varianten auch darauf ab, ähnliches Material etwa über den automatisierten Abgleich der biometrischen Daten von Tätern und Opfern oder der in einer Datei enthaltenen Metadaten (Zeit der Aufnahme, genutztes Aufnahmegerät, Ortsdaten) mit aus anderen Darstellungen bekannten Personen und Orten zu identifizieren. Die Effektivität dieser Ansätze hängt von dem Umfang der bereits identifizierten CSA-Darstellungen ab; hier arbeiten die Ermittlungsbehörden der EU-Mitgliedstaaten mit einer gemeinsamen bei Interpol angesiedelten supranationalen und entsprechend umfangreichen Datenbank (International Child Sexual Exploitation [ICSE] database). Die großen Plattformen arbeiten in diesem Bereich ebenfalls mit automatisierten Erkennungssystemen. Im Bereich von terroristischen Propagandainhalten (z. B. Enthauptungsvideos) hat sich eine gemeinsame Datenbank von Onlineplattformen mit Hashwerten identifizierter Darstellungen etabliert („database of hashes"). Bei dem Upload einer geblacklisteten Datei durch einen Nutzer sperren die Anbieter diesen Inhalt unverzüglich.

Mit der Identifikation neuer Darstellungen haben beide Verfahren Probleme. Sie sind auf den Anwendungsbereich der Identifizierung bekannter Darstellungen optimiert und können so zeitaufwendige – und gegebenenfalls auch psychisch belastende – inhaltliche Analysen durch Menschen insbesondere bei großen Dateimengen zuverlässig ersetzen.[2]

Automatisierte Inhalteerkennung und Inhaltemoderation

Weitergehende maschinenlernbasierte Ansätze in der Bilderkennung zielen auf Formen automatisierter Inhalteerkennung ab. Die Identifikation bestimmter Objekte, Personen oder Handlungen ist dabei ein klassisches Anwendungsfeld maschinellen Lernens. Hier existiert eine Vielzahl unterschiedlicher Ansätze mit sehr unterschiedlichen Kombinationen neuronaler Netzwerke und einer ganzen Reihe bereits trainierter Modelle, die Aussagen zu den dargestellten Inhalten (Objekterkennung) und Handlungen (Aktionserkennung) von neuen Darstellungen produzieren. Die Erfolgsquoten sind dabei je nach Anwendungsdomäne unterschiedlich hoch – die Klassifikationsaufgaben funktionieren nicht bei allen jugendschutzrelevanten Darstellungen gleich gut. Während die Erkennung von Nacktheit, Drogenkonsum und einzelnen Gewaltformen (Waffengewalt, Explosionen, Blut, Selbstverletzung, Unterernährung) in audiovisuellen Medien relativ gut funktioniert, ist die automatische Identifikation von Fake News, ängstigenden Inhalten oder Massenpaniken eher schwierig.

Auch für Textformate werden automatisierte Erkennungsverfahren benutzt, die klassischerweise auf bestimmte Stichworte achten, diese in automatisierten Heuristiken bewerten[3], in komplexeren Varianten aber auch die Nähe von einzelnen Wörtern zu anderen Wörtern erkennen (sogenannte Bag-of-Words-Verfahren), neue sprachliche Codes lernen können und bereits seit Längerem die Einstellung und vermeintliche Stimmung des Äußernden in die Bewertung mit einbeziehen können („sentiment analysis"). Dort, wo derartige Textanalysen differenzierte Bewertungsergebnisse erlauben, werden sie auch auf die Ton- bzw. Sprachspur audiovisueller Inhalte angewandt, nachdem diese – wiederum mit aktuellen KI-Verfahren („natural language processing" [NLP]) – automatisiert transkribiert und auf diese Weise *verstanden* worden sind. Daneben werden auch Verfahren eingesetzt, die bestimmte Formen des Meldeverhaltens durch Nutzerinnen und Nutzer über Zeit analysieren und damit Rückschlüsse auf problematische Darstellungen gewinnen.

Insgesamt führt künstliche Intelligenz vor allem dazu, bestimmte Inhalte zu identifizieren – wie Nacktheit, Verletzungen oder toxische (Hass-) Rede. Gelten entsprechende Inhalte im jeweiligen Angebot als absolut verboten, vor allem über die Nutzungsbedingungen oder Community Standards, so können die derzeitigen Verfahren relativ zuverlässige automatisierte Entscheidungen treffen, d. h. die entsprechenden Inhalte löschen. Das vor allem methodische Problem dabei ist, dass angesichts ausbleibender Nachkontrollen und Rückmeldungen durch Löschungsbetroffene stets unklar bleibt, wie viele gelöschte Inhalte falsch zugeordnet und so zu Unrecht gelöscht wurden. Es gibt keine systematischen Nachkontrollen und die vorgehaltenen Beschwerdeverfahren sind vielfach aufwendig oder unbekannt. Zahllose Darstellungen sind in ihrem jeweiligen Kontext gegebenenfalls gerade keine Verletzung von Community-Vorschriften, werden aber vermeintlich als solche erkannt, weil die KI keine oder allenfalls eine rudimentäre Kontexterkennung in ihre mathematische Wahrscheinlichkeitsberechnung einbe-

zieht. Angesichts der Unsicherheiten bei der automatisierten Erkennung setzen etliche Anbieter KI-Verfahren vor allem für die Identifizierung und Vorauswahl möglicher jugendschutzrelevanter Inhalte ein. Es sind am Ende Menschen, die diese vorausgewählten Inhalte sichten und dann auf mögliche Rechtsverstöße hin untersuchen (sogenannte entscheidungsunterstützende KI oder „decision support systems" [DSS]).

Assistierte und automatisierte Altersbewertungen

Einen Schritt weiter gehen technische Ansätze, die nicht nur Inhalte und Themen einer Darstellung analysieren und erkennen, sondern auf dieser Grundlage Aussagen über die mögliche altersbezogene Jugendschutzrelevanz treffen. Derartige Entscheidungen zu automatisieren, ist alles andere als trivial, da in praktisch allen regionalen und nationalen Jugendschutzrahmen eine solche Altersbewertung im jeweiligen kulturellen Kontext erfolgt – und der kann sehr unterschiedlich sein, wie auch die Filmfreigaben in der Vergleichsrubrik der *tv diskurs* zeigen.

Vor allem mit Blick auf die mangelnde Kontexterkennung von automatisierten Formen der Inhaltserkennung ist derzeit noch davon auszugehen, dass Verfahren maschinellen Lernens Inhalte erkennen, aber deren Entwicklungsbeeinträchtigung in der Gesamtschau gerade nicht. Die Pilotversuche von jugendschutz.net sind in technischer Hinsicht insoweit zu begrüßen, ihre Analyseergebnisse aber werfen mehr Fragen als Antworten auf. Die hohen[4] Erkennungsraten bei Selbstverletzungen und Enthauptungsvideos sprechen für die mittlerweile recht gut arbeitenden Algorithmen und Modelle bei der Inhaltserkennung, sie sagen aber nichts über die Rechtmäßigkeit der identifizierten Darstellungen allein und mit Blick auf den Kontext ihrer Veröffentlichung aus.[5] Selbstverletzungsdarstellungen können im Rahmen von Informationsangeboten oder als Warnungen vor der Nachahmung veröffentlicht werden, Enthauptungsvideos können als Dokumente gravierender Menschenrechtsverletzungen eine Daseinsberechtigung haben und gegebenenfalls in den Anwendungsbereich von § 5 Abs. 6 JMStV fallen. Außerdem sieht das deutsche Jugendmedienschutzrecht differenzierte Altersgruppen einer möglichen Entwicklungsbeeinträchtigung vor (0 – 5, 6 – 11, 12 – 15, 16 – 17 Jahre). Diese Form der altersgruppenbezogenen Bewertung durch die Analyse der audiovisuellen Inhalte ist jedenfalls für derzeitige Formen künstlicher Intelligenz ausgeschlossen. Kurzum: Eine automatisierte Inhaltserkennung ermöglicht in den meisten Fällen keine automatisierte jugendschutzrechtliche Altersbewertung. Alle anerkannten Einrichtungen der freiwilligen Selbstkontrolle (FSF, FSM, FSK.online, USK.online) arbeiten oder planen daher in erster Linie mit Assistenzsystemen bei der gestützten Altersbewertung von Einzelinhalten.[6] Eine deutliche Verbesserung der derzeit unzureichend mit Metadaten versehenen Datengrundlage wird hier möglicherweise die Umsetzung der AVMD-Richtlinie bis Herbst 2020 bringen, die eine ganze Reihe technikbezogener Auflagen insbesondere für Video-Sharing-Anbieter wie YouTube und Vimeo macht.

Präventive KI: Erkennung problematischer Interaktion und Kommunikation auf Endgeräteebene

Interessante und weiterführende Möglichkeiten bieten Formen künstlicher Intelligenz im Rahmen von interaktiver Onlinekommunikation: So können z. B. anhand spezifischer, typischer Bullying- oder auch Grooming-Verläufe jugendschutzrelevante Verhaltensweisen zwischen Kommunikationspartnern analysiert und identifiziert werden. Die KI-Forschung ist hier noch am Anfang, aber vielversprechend: Auf der Ebene des Endgeräts eingesetzt, ermöglichen entsprechende Verfahren Formen der Echtzeit-Überwachung und -Rückhaltung von Kommunikationen, die sowohl zu Kindern als Rezipienten gelangen, als auch von diesen selbst produziert und gegebenenfalls ausgesendet werden („contextual safety", z. B. bei Bullying, Grooming, Rachepornos oder toxischer [Hass-]Rede). Entsprechende Assistenzsysteme können produzierende oder aktive Nutzerinnen und Nutzer vom Versenden abhalten, bei passiven Nutzern Aufmerksamkeit für die mögliche (auch rechtliche) Relevanz schaffen und potenzielle Risiken erklären. In diese Richtung gehen derzeit in der Entwicklung befindliche Assistenzsysteme („cyber safety digital assistants", „online safety assistants") wie OYOTY[7] für Kinder und Jugendliche oder bark[8] für Eltern, die Warnhinweise bei riskantem Verhalten ihrer Kinder erhalten.

KI im Kinderalltag: Assistenten als neue Vermittler und Begleiter

Jugendschutzbezogene Assistenzsysteme sind von Ausnahmen abgesehen derzeit noch eher Vision als Realität. Ganz real dagegen ist die derzeitige Nutzung von bestehenden Assistenzsystemen auf mobilen Endgeräten und Smart Speakern mit digitalen Sprachassistenten (Alexa, Google Home, Siri, Cortana etc.) im neuerdings schlauen Zuhause (Smart Home). Kinder und Jugendliche interagieren mit diesen KI-gestützten[9] Systemen zunehmend häufig.[10] Die Kritik einer von jugendschutz.net durchge-

führten Analyse zielte hier auf die mangelnde Berücksichtigung unterschiedlicher altersgerechter Inhalte ab und – was noch schwerer wiegt – auf die mangelnden Konfigurationsmöglichkeiten für Minderjährige und insbesondere sehr junger Nutzerinnen und Nutzer entsprechender Systeme. So gewähren die ratlosen Assistenten teils Zugriff auch auf jugendschutzrelevante Inhalte, bei Assistenten mit Displays etwa auf entwicklungsbeeinträchtigende oder gar jugendgefährdende Darstellungen.[11] Hier scheint das Prinzip von Safety by Design noch nicht bei allen Unternehmen in der Produktentwicklung angekommen zu sein.[12]

Die Kehrseite der Medaille von KI-Systemen, mit denen Kinder und Jugendliche interagieren, ist, dass die Nutzungsdaten derzeit regelmäßig nicht lokal auf dem Endgerät, sondern auf Servern in der Cloud des jeweiligen Anbieters verarbeitet und gespeichert werden – nicht nur, aber auch zur Verbesserung der lernenden Systeme.[13] Damit strukturell einher gehen Formen der ständigen Überwachung, der biometrischen Auswertung und des Profilings – mit gegebenenfalls der Autonomie abträglichen Folgen, da allein die Kenntnis der Überwachung bei den Überwachten schon zu (vermutet) sozialadäquatem Verhalten führen kann.

Anmerkungen:

1 Jugendmedienschutz als rezipientenbezogener Schutzzweck tritt hier in den Hintergrund; Hauptzweck der Strafvorschriften im Bereich sexuellen Kindesmissbrauchs ist der Schutz der sexuellen Selbstbestimmung von Kindern und damit der Opferschutz; Renzikowski, MüKo, 3. A., 2017, § 176 StGB Rn. 3

2 Diesem Ansatz folgt auch das Projekt „NRW-Ansatz"; vgl. auch die Erfahrungen mit Vorauswahlverfahren für CSA-Material bei Google, abrufbar unter: https://www.standard.co.uk. Vgl. auch die Erfahrungen mit dem Einsatz entsprechender Verfahren bei Facebook, abrufbar unter: https://newsroom.fb.com

3 Diesen Ansatz verfolgen z. B. JusProg und PureSight bei der automatisierten Altersbewertung bislang unbekannter Internetangebote.

4 Bei Selbstverletzungen spricht jugendschutz.net von 95 % Erkennungsrate, wobei weder klar wird, wie das Testbed zusammengesetzt war, noch wie viele „false positives" und „false negatives" unter den verbleibenden 5 % waren. Unklar blieb auch, mit welcher Genauigkeit diese Erkennung stattgefunden hat. Die „accuracy" beschreibt die Wahrscheinlichkeit, mit der ein Algorithmus eine Klassifikationsentscheidung getroffen hat.

5 Ein gutes Beispiel dafür sind die Jugendschutzfilter von Tumblr, die auch jene Darstellungen als unzulässig markierten, die der Anbieter selbst als Beispiele zulässiger Nacktheit vorhielt, etwa aus dem Bereich der klassischen Kunst. Vgl.: Hungo, H.: *Tumblr's Porn Filter Flags Its Own Examples of ‚Permitted' Nudity*. In: Gizmodo, 17.12.2018. Abrufbar unter: https://gizmodo.com

6 IARC für Onlinegames und Apps seit 2013 (checklistenbasiert, mit lokalisierten Altersfreigaben); angekündigter FSK-Pilot (checklistenbasiert); YouRateIt (kurze Checkliste für nutzergenerierte Inhalte). Die KI besteht hier vor allem in der Nutzung von Algorithmen mit determiniertem Ausgang, d. h., die Checklisten werfen in Wenn-Dann-Relationen bestimmte Altersbewertungen aus. Weiter geht Netflix, das auf Grundlage umfangreicher Verschlagwortung der eigenen Inhalte und dem Wissen über bestehende Altersklassifikationen weltweit automatisiert und halb automatisierte Altersbewertungen und lokale Altersanpassungen vornimmt; das Unternehmen führt derzeit Pilotprojekte in Australien und Großbritannien durch, wo Netflix selbst anstatt der vorgeschriebenen Rating Boards die Klassifizierung vornimmt. Vgl.: Loussikian, K.: *Netflix gets approval to classify own shows after two-year trial*. In: The Sydney Morning Herald, 20.01.2019. Abrufbar unter: https://www.smh.com.au. Vgl. auch: Waterson, J.: *Netflix to set its own age ratings for film and television programmes. British Board of Film Classification allows streaming giant to rate content*. In: The Guardian, 14.03.2019. Abrufbar unter: https://www.theguardian.com

7 Abrufbar unter: https://www.oyoty.com/ (beschränkt auf Instagram, Facebook und Twitter)

8 Abrufbar unter: https://www.bark.us/

9 Die Assistenten nutzen NLP-Verfahren („natural language processing") zur Erkennung der sprachlichen Befehlseingabe und setzen den so erkannten Befehl softwareseitig – wiederum oft mithilfe von KI – um.

10 Im Rahmen der *KIM-Studie 2018* verfügten 6 % der befragten Haushalte mit jüngeren Kindern über Sprachassistenten (S. 8), die *JIM-Studie 2018* erwähnt diese Geräte in 12 % der Haushalte mit Kindern zwischen 12 und 19 Jahren (S. 6).

11 Vgl.: jugendschutz.net: *Praxis-Info: „Alexa, was hältst du von Jugendschutz?" Sprachassistenzsysteme noch nicht auf Nutzung von Familien ausgelegt*. Mainz 2018, S. 5. Abrufbar unter: https://www.jugendschutz.net

12 Ein guter Überblick über die kindersichere Konfiguration von Sprachassistenten ist abrufbar unter: https://www.surfen-ohne-risiko.net

13 Vgl. die Information zu Sprachassistenten bei Annas Leben, abrufbar unter: https://www.annasleben.de

Dr. Stephan Dreyer ist Senior Researcher für Medienrecht und Media Governance am Leibniz-Institut für Medienforschung | Hans-Bredow-Institut (HBI).