Institut für Kommunikationsakustik

humAIne
HUMAN CENTERED AI NETWORK

HUMAINE-TOOLBOX

INSTRUMENTE ZUR HUMANZENTRIERTEN TECHNIKENTWICKLUNG

VORGEHENSMODELL ZUR HUMANZENTRIERTEN EINFÜHRUNG VON ACTIVE LEARNING

ERKENNTNISZIEL
Das Dokument gibt einen Überblick über Fragestellungen, die beim Einsatz von Active Learning auftreten können.

EINORDNUNG IN DAS HUMAINE-METHODENSPEKTRUM
Anforderungsanalyse

AUTOREN
Alfredo Virgillito, Anja Gerlmaier, Dorothea Kolossa, Jan Freiwald, Marie Ossenkopf, Max Bauroth, Pavlos Rath-Manakidis, Peter Sertdal, Sophie Berretta

ANSPRECHPARTNER
Jan Freiwald (jan.freiwald@rub.de)

STAND
Mai 2022

GEFÖRDERT VOM
Bundesministerium für Bildung und Forschung

Kompetenzzentren Arbeitsforschung

https://humaine.info/

	KONKRETE BESCHREIBUNG
Erkenntnisziel beim Einsatz des Instrumentes	Das Dokument gibt einen Überblick über Fragestellungen, die beim Einsatz von Active Learning auftreten können.
Welche Fragestellungen können untersucht werden?	Was ist Active Learning? Welche humanzentrierten Probleme werden durch AL behandelt? Welche Aspekte der Arbeitsregulation müssen beachtet werden? Was muss beim Datenschutz beachtet werden? Wie sieht es mit dem Arbeits- und Gesundheitsschutz aus? Über Arbeitsgestaltungsanforderungen beim AL. Welche Rolle spielt Flow? Wie setzt sich der Konfektionierungsaufwand zusammen? Wie sieht der Trainingsprozess aus? Was ist Explainability und warum ist das sinnvoll? Wer ist verantwortlich? Diskriminierungsfreiheit und Fairness.
Zu erwartende Ergebnisse	Der Leser hat einen Überblick über Active Learning.
Typische Anlässe für den Einsatz	Der Leser möchte Active Learning einsetzen, oder ist generell interessiert.
Einordnung in das Spektrum der Untersuchungsmethoden	Anforderungsanalyse
Welche Kenntnisse werden für den Einsatz des Instrumentes benötigt?	Vorkenntnisse in Machine Learning Verfahren sind erforderlich. Programmiererfahrung ist wünschenswert.
Wie viele Personen und welcher Gesamtzeitaufwand werden für den Einsatz des Instruments benötigt? (Erhebung und Auswertung)	n.a
Welcher Zeitaufwand wird auf Seiten des Unter-suchungspartners benötigt?	45 min Lesezeit + Diskussion
Besonderer Nutzen / Empfehlung zum Einsatz	n.a.
Empfohlene Zitation des Instruments	n.a
Zu beachtendes Copyright	Alle Rechte vorbehalten.
Literaturverweise und/oder andere Referenzen zum Einsatz des Instrumentes	Literatur ist im Text vermerkt.
Kontakt/Ansprechpartner	Jan Freiwald (jan.freiwald@rub.de)

WICHTIGE HINWEISE

Das HUMAINE Forschungs- und Entwicklungsprojekt wird / wurde durch das Bundesministerium für Bildung und Forschung (BMBF) im Programm „Innovationen für die Produktion, Dienstleistung und Arbeit von morgen“ gefördert und vom Projektträger Karlsruhe (PTKA) betreut. Die Verantwortung für den Inhalt dieser Veröffentlichung liegt bei den Autor:innen[1]. Dieses Dokument ist, laut Vorhabensbeschreibung, öffentlich verfügbar.

ANSPRECHPARTNER:INNEN

Aktuelle Ansprechpartner finden Sie auf der Projekt-Homepage (https://humaine.info/). Bei Kritik, Ergänzungs- oder Änderungswünschen wenden Sie sich bitte zunächst an den korrespondierenden Autor oder die Projektverantworlichen.

Korrespondenz Jan Freiwald

Autor:innen (alphabetisch) Alfredo Virgillito, Anja Gerlmaier, Dorothea Kolossa, Jan Freiwald, Marie Ossenkopf, Max Bauroth, Pavlos RathManakidis, Peter Sertdal, Sophie Berretta

Editor:innen Jan Freiwald

VERSIONSHISTORIE

31.12.2021 Veröffentlichung des ersten Entwurfs. Mit diesem Dokument liegt eine umfangreiche Anforderungsanalyse vor. Das Dokument unterliegt noch einem Änderungsprozess und ist nicht final.

[1] *Alle Informationen in diesem Dokument dienen der allgemeinen Information. Sie stellen keine Rechtsberatung im Einzelfall dar, können und sollen diese auch nicht ersetzen. Die dargestellten Inhalte werden mit größtmöglicher Sorgfalt zusammengestellt. Dennoch müssen wir die Haftung für die Vollständigkeit, Richtigkeit und Aktualität der eigenen Informationen, die in diesem Dokument zur Nutzung bereitgehalten werden, ausschließen, es sei denn, wir handeln vorsätzlich oder grobfahrlässig. Der Haftungsausschluss gilt auch für Linksammlungen und Quellen, die zurzeit Bestehen oder in Zukunft bestehen werden.*

1 EINLEITUNG

Dieses Dokument soll als Vorgehensmodell für die humanzentierte Einführung und Verwendung von Active Learning (AL) im Business-Kontext dienen. Es ist ein lebendes Dokument, das heißt, dass es während der Projektlaufzeit erweitert und verbessert wird. Mit diesem Dokument wird ein Fragenkatalog zur Verfügung gestellt, an Hand dessen Active Learning im Unternehmen spezifiziert und human-zentriert eingeführt werden kann. Die Inhalte, die in diesem Dokument behandelt werden, wurden von einem interdisziplinären Team innerhalb der HUMAINE Projekts erarbeitet und ausgestaltet. Es werden zunächst allgemeine Sachverhalte zum Thema AL dargestellt; darauf aufbauend werden human-spezifische Aspekte im Kontext des AL näher beleuchtet.

1.1 WAS IST ACTIVE LEARNING?

Die Leistungsfähigkeit moderner Datenverarbeitung im Bereich des überwachten Lernens beruht, unter anderem, auf der Erschließung und Verwendung großer Datenmengen. Vollständig überwachte Lernalgorithmen benötigen, zusätzlich zu dieser großen Menge von Eingangsdaten, noch ein sogenanntes Label für jeden Datenpunkt. Ein Label beschreibt die Grundwahrheit, die der Algorithmus über den entsprechenden Datenpunkt lernen soll, also die Information, die aus den Eingangsdaten ermittelt werden soll. In vielen Anwendungen ist das Erstellen dieser Label eine aufwendige und teure Aufgabe, die oft von Domänen-Expert:innen übernommen werden muss. Die Idee des AL besteht darin, einem lernenden System die Möglichkeit zu geben, durch geschicktes Auswählen von Datenpunkten aus der Gesamtmenge an Daten, die Anzahl an benötigten Labeln zu reduzieren. Dadurch werden die Annotator:innen entlastet und zusätzlich kann die Konvergenzgeschwindigkeit der Lernalgorithmen erhöht werden, wodurch mit weniger gelabelten Trainingsdaten bereits eine gute Performanz erzielt werden kann. In (Settles 2009) findet sich eine umfangreiche und detailierte Beschreibung der technischen Aspekte des AL. In den letzten Jahren hat sich die AL-Forschung immer mehr in Richtung von Gradienten-basierten Verfahren, neuronalen Netzen entwickelt, dazu finden sich einige Untersuchungen, z.B. in (Chakraborty, Balasubramanian und Panchanathan 2015; Singh und Chakraborty 2020). Die Entwicklung und Anwendung solcher Verfahren ist ein aktuelles Forschungsthema.

1.2 WELCHE HUMANZENTRIERTEN PROBLEME WERDEN DURCH AL BEHANDELT?

Im humanzentrierten Kontext ist die Antwort auf diese Frage zweigeteilt:

- Erstens kann beim Active Learning, laut (Settles 2009), durch gezieltes Auswählen von geeigneten Trainingsbeispielen ein maschineller Lernalgorithmus möglichst effizient trainiert werden. Dadurch wird die Arbeitszeit reduziert, die mit dem Erstellen von Labeln verbracht werden muss.
- Zweitens kann vermutlich durch eine Zusammenarbeit von Mensch und Maschine eine schnellere Annotation und eine höhere Labelqualität erzielt werden. Die Maschine kann dem Menschen z.B. ein Label vorschlagen, auf Details hinweisen, oder auf vermeintliche

Fehler aufmerksam machen, wenn die menschliche und maschinelle Einschätzung auseinander gehen. Zu diesem Punkt ist Forschung notwendig.

2 ANFORDERUNGSANALYSE

In diesem Kapitel finden sich Fragen zur Anforderungsanalyse.

2.1 WELCHE ASPEKTE DER ARBEITSREGULATION MÜSSEN BEACHTET WERDEN?

Falls es einen Betriebsrat gibt, sollte mit diesem eine Einführung von Active Learning besprochen werden. Das gilt vor allem, wenn viele Mitarbeiter von der Einführung betroffen sind. Betriebsräte dürfen dazu eigene Berater:innen zu Rate ziehen und somit ihre Expertise erweitern (§ 80 (3) BetrVG), um informiert entscheiden zu können. Generell sind darüber hinaus Betriebsräte auf Basis des Betriebsverfassungsgesetzes bei der Gestaltung von Arbeitsplätzen hinzuzuziehen. Es ist zu klären, ob das Verfahren sinnvoll einsetzbar ist und wie Beschäftigte an das Verfahren herangeführt und bei der Einführung beteiligt werden. Dazu sind verschiedene Modelle denkbar, z.B. ein Test des Verfahrens mit einer begrenzten Zahl von freiwilligen Mitarbeitenden. Sinnvoll ist die frühe Einbindung der Mitarbeitenden. In Betrieben ohne Betriebsrat ist die Nutzung einer alternativen Beteiligungsform empfehlenswert, z.B. Runde Tische. Wichtig ist dabei, dass AL tatsächlich eine Arbeitserleichterung darstellt. Es ist also zu prüfen, ob nach der Einführung eines AL-Verfahrens eine Arbeitsentlastung stattgefunden hat, z.B. durch eine Nutzendenumfrage. Zusätzlich kann der Einsatz von AL auch nutzenstiftend im Sinne der Förderung von Motivation oder Vigilanz der Arbeitenden sein und somit zur Arbeitsqualität beisteuern.

2.2 WAS MUSS BEIM DATENSCHUTZ BEACHTET WERDEN?

Die wichtigsten Punkte zum Datenschutz werden in der DSGVO (Verordnung (EU) 2016/679 2016) geregelt. Hier ist unter anderem zu beachten, dass personenbezogene Daten nur zweckgebungen gesammelt werden dürfen. Darüber hinaus ist ein Datenschutzkonzept im Interesse der Beschäftigten zu entwickeln und tranzparent zu machen. Damit soll eine Überwachung der Beschäftigten, speziell eine Leistungsüberwachung verhindert werden, welche auch generell mitbestimmungspflichtig sind (§75 (2) & §87 BetrVG). Je nach Datensatz kann es sinnvoll sein, den Beschäftigten oder Vertretenden Vollzugriff auf die erhobenen Daten zu gewähren, um das Vertrauen zu steigern. Wenn beim Labeln Metadaten über die Beschäftigten erhoben werden müssen, dann muss auch eine Mitbestimmung der Beschäftigten bei der Auswertung der Daten gewährleistet sein. Das AL-System ist abschaltbar zu gestalten und für die Daten muss ein Löschkonzept vorliegen. Je nach Arbeitsstätte ist zu klären, welche Informationen den Nutzer:innen des AL-Systems angezeigt werden. So kann es z.B. bei mobiler Arbeit, beim Out-/Crowdsourcing, bei Co-Working-Spaces o.Ä. notwendig sein, bestimmte Informationen nicht anzuzeigen, z.B.

personenbezogene Daten, wie Patientennamen, wenn es sich um einen medizinischen Datensatz handelt.

2.3 WIE SIEHT ES MIT DEM ARBEITS- UND GESUNDHEITSSCHUTZ AUS?

Obwohl durch AL Menschen entlastet werden können, ist es notwendig, dadurch entstehende neue Probleme proaktiv zu behandeln. Da AL vor allem in Bildschirmarbeit stattfindet, ist es sinnvoll, regelmäßige Mikropausen (ggf. mit Bewegungsangebot) in die Tätigkeitsausführung einzuplanen (Fleischer u. a. 2013; Hüttges, Müller und P. Richter 2005; Kim, Park und Headrick 2018). Falls das Labeln der Daten emotional fordernd ist (z.B. bei Anwendungen, die Gewaltdarstellungen oder Unfälle analysieren), sollte die Nutzerschnittstelle Kontaktdaten zu Beratungsangeboten bereitstellen (Steiger u. a. 2021). Es gibt aber auch niederschwelligere Ideen, die Arbeit systematisch netter zu gestalten – in der forensischen Linguistik läßt z.B. Prof. Tatjana Scheffler die Annotator:innen von Hassnachrichten zwischendurch (verpflichtend) niedliche InternetInhalte eigener Wahl, wie z.B. Katzenvideos, anschauen. Darüber hinaus sollten KI-Trainer:innen / Annotator:innen in der Lage sein, das Labeln eines Datenpunktes abzulehnen oder auf später zu verschieben, wenn es zu belastend ist. Generell sollten psychischer Stress und Verdichtungsmomente präventiv eingedämmt werden. Gesundheitlichen Beeinträchtigungen der Nutzer:innen müssen durch eine Gefährdungsbeurteilung psychischer Belastung bzw. Technologiefolgeabschätzung verhindert werden. Festgestellte Probleme sollten im Betriebsrat / am Runden Tisch diskutiert werden. Dies kann, falls notwendig, bis zur Abschaltung des Systems eskaliert werden.

2.4 Über ARBEITSGESTALTUNGSANFORDERUNGEN BEIM AL

Beim AL können für die Annotatierenden verschiedene Formen von psychischen Fehlbeanspruchungen auftreten, die sich sowohl ungünstig auf das Arbeitsergebnis, als auch auf das psychische Wohlbefinden auswirken (Berka u. a. 2007; Warm, Matthews und Finomore Jr 2008). Infolge der teilweise sehr einseitigen kognitiven Anforderungen bei der Tätigkeit des Labelns (Daueraufmerksamkeit) ergibt sich das Risiko von verminderter Vigilanzleistungen sowie Monotonieerleben. Die wiederholte Darstellung von gleichartigem Bildmaterial begünstigt eine Unterforderung der intellektuellen und Überforderung der sinnlichen Leistungsvoraussetzungen der Annotator:innen. Dieser erlebt Langeweile und Unlustgefühle, die Aufgaben fortzusetzen. Das Erleben von Langeweile wie auch erhöhter Vigilanz wirkt sich im Verlauf der Tätigkeit potenziell ungünstig auf die Arbeitsmotivation der Annotator:innen aus und erhöht in erheblichem Maße die Wahrscheinlichkeit von Fehlern durch eine Abnahme der Konzentration (Warm, Matthews und Finomore Jr 2008). Eine weitere Form von Fehlbeanspruchung stellt psychische Sättigung dar. Dies kann in Fällen auftreten, bei denen die Annotierenden in zunehmendem Maße der Bearbeitung der Aufgabe überdrüssig werden und die Sinnhaftigkeit derselben verloren geht. Statt Langeweile können die Annotierenden trotz einer generellen Bereitschaft zur Aufgabenrealisierung eine unlustbetonte Gereiztheit und Widerwillen empfinden (Peter Richter und Hacker 2000). Diese Form der Fehlbeanspruchung kann zu Ärgerreaktionen führen und kontraproduktives

Verhalten nach sich ziehen, z.B. bewusste fehlerhafte Eingaben (Bessiere u. a. 2004). Zusätzlich kann es durch die dauerhafte Sichtung von belastendem Material z.B. Gewaltdarstellungen zu physiologischen Stressreaktionen bei Annotator:innen kommen. Die Annotierenden fühlen sich durch das dargestellte Material emotional überfordert und erleben Spannungszustände oder Ängste, die auch noch längere Zeit nach der Ausübung der Tätigkeit auftreten und bis hin zu physiologischen Reaktionen wie Herzrasen, Ängsten und depressiven Verstimmungen führen (Steiger u. a. 2021). Die beschriebenen ungünstigen Beanspruchungsfolgen und Auswirkungen auf das Leistungsverhalten zeigen einen erheblichen Gestaltungsbedarf bei Annotationstätigkeiten auf. Hier ist es zum einen sinnvoll, die Dauer der Annotation pro Tag zeitlich zu beschränken. Ein anderer Ansatz besteht darin, gezielte Tätigkeitswechsel einzuplanen. Hierbei können beispielsweise beim Labeln von Bildmaterial systematisch andere z.B. kommunikative oder motorische Betätigungen (z.B. Dokumente vom Gemeinschaftsdrucker abholen o. ä.) für einen Ausgleich sorgen.

2.5 WELCHE ROLLE SPIELT FLOW?

Flow bezeichnet einen als angenehm empfundenen mentalen Erfahrungszustand, der sich durch Konzentration und Absorption einer Person während einer optimal herausfordernden Tätigkeit zeigt (Csikszentmihalyi 1975). Eine optimal herausfordernde Tätigkeit zeichnet sich dadurch aus, dass sich die Fähigkeiten einer Person und die Anforderungen der Aufgabe im Gleichgewicht, leicht über dem Durchschnitt, befinden (Csikszentmihalyi und LeFevre 1989). Während des Flow-Erlebens zeigen Personen typischerweise ein verzerrtes Zeitempfinden sowie die Verschmelzung von Bewusstsein und Handeln, sodass Reize außerhalb der Tätigkeit ausgeblendet werden (Csikszentmihalyi 1975). Durch das Erleben von Flow wird die Leistung, das Wohlbefinden, die Einstellung als auch das soziale Umfeld einer Person positiv beeinflusst (Peifer und Wolters 2017). Durch die positiven Effekte von Flow ist essowohl aus Sicht der Organisation als auch aus Sicht der Arbeitnehmenden von großem Interesse das FlowErleben während alltäglicher Tätigkeiten zu fördern (Peifer und Wolters 2017). Um in den Flow-Zustand zu gelangen, wurden bereits verschiedene Flow-förderliche Faktoren identifiziert. Eine besondere Rolle bei der Entstehung von Flow spielen die Facetten der Aufgabengestaltung, wie Autonomie (Emanuel, Colombo und Zito 2016), die erlebte Wichtigkeit der Aufgabe (Bassi und Fave 2012) und Feedback (Maeran und Cangiano 2013). Ebenfalls entscheidend für die Entstehung des Flow-Zustandes ist eine Balance zwischen der Anforderung der Aufgabe und der Fähigkeit der Person (Bakker 2005). Auch das soziale Umfeld, wie die Unterstützung durch Kolleg:innen (Bakker 2008) und das Unternehmensklima (Fagerlind u. a. 2013) wurden als Flow-förderliche Faktoren identifiziert. Bei Tätigkeiten, die durch Arbeit mit einem ALSystem entstehen, kann davon ausgegangen werden, dass diese ohne weiteres Zutun wenig Flow-förderliche Faktoren beinhaltet, da es sich eher um monotone Tätigkeiten mit geringer kognitiver Anstrengung handelt. Dieses steht im Widerspruch zu den Voraussetzungen, die für das Flow-Erleben identifiziert wurden. Durch die positiven Effekte auf die Leistung und das Wohlbefinden einer Person (Peifer und Wolters 2017) ist es wünschenswert Flow-Erleben auch beim AL zu ermöglichen. Aus anderen Kontexten mit ähnlichen monotonen Aufgaben (z.B. aus dem Schaltanlagenbau) zeigt sich, dass das Flow-Erleben durch hohe Entscheidungs-, Methoden- und Planungsautonomie

positiv beeinflusst werden kann (Tausch und Peifer 2019). Beim AL könnte dies durch z.B. eine autonome Pausengestaltung während der Annotation realisiert werden. Weitere Hinweise zur Förderung des Flow-Erlebens in monotonen Aufgaben finden sich in der Untersuchung von Sillatos (2014), indem durch Gamification und Immersion das Flow-Erleben gefördert wurde. Auch das Einbauen von positivem Feedback kann zur Steigerung des Flow Empfindens führen (Maeran und Cangiano 2013). Inwiefern die Steigerung des Flow-Erlebens bei Aktivitäten des AL realisierbar wären gilt es noch zu untersuchen. Bisher gibt es wenig explizite Untersuchungen zu Flow beim AL, daher sind hier auch noch viele Möglichkeiten für aktuelle und menschenzentrierte Forschung.

2.6 WIE SETZT SICH DER KONFEKTIONIERUNGSAUFWAND ZUSAMMEN?

AL erfordert Hardware, die Machine-Learning-fähig ist. Dazu gehören unter anderem Festplattenspeicher, der alle Daten halten kann, eine oder mehrere starke CPU/GPU (je nach Algorithmus), Netzwerkinfrastruktur und Backupsysteme. Es kann hilfreich sein, bestehende Daten zu nutzen. Beim Active Learning ist ein vortrainiertes Modell hilfreich. Dazu gibt es in vielen Anwendungsfällen freie Datenbanken, die zumindest zum Vortraining eines MLAlgorithmus verwendet werden können. Falls es im speziellen Anwendungsfall keine freien Datenbanken gibt, oder diese durch Lizenzen zu sehr eingeschränkt sind, müssen initiale Daten gekauft werden; ggf. muss ein initiales Modell gekauft werden. Wenn nur eigene Daten vorliegen, sollte ein initialer Anteil bereits gelabelt sein, um schneller in einen effizienten Arbeitsbereich des AL zu kommen. Es kann notwendig sein, Annotator:innen zu bezahlen oder anders zu vergüten. Hinzu kommen Kosten, die dadurch entstehen, dass die Labels u.U. von niedriger Qualität sind, wenn sie nicht durch Experten erstellt werden. Das ist vor allem beim Outsourcing/Crowdsourcing zu beachten. Zusätzlich muss beachtet werden, dass ein Supportaufwand anfällt, wenn Kunden/Nutzer im Self-Service Labeling und AL betreiben sollen. Daher ist es je nachdem, wie das Active Learning eingeführt wird, notwendig, die Nutzerschnittstelle zu implementieren, zu supporten und zu warten. Wenn das gesamte AL-System im Unternehmen entwickelt wird, ist neben Domänenwissen eine hohe KI-Expertise notwendig. Durch die Verwendung von komplexen Systemen innerhalb des Unternehmens können Fortbildungskosten entstehen, wenn Mitarbeiter:innen an das System herangeführt werden müssen. Je nach Aufgabenstellung stellt automatisches Machine Learning (AutoML) eine Lösung dieses Problems dar, da weniger Expertise notwendig ist und bei der Implementierung weniger Fehler gemacht werden.

2.7 WIE SIEHT DER TRAININGSPROZESS AUS?

Beim Active Learning wird der Mensch direkt in den Trainingsprozess eingbunden, das sogenannte Humanin-the-Loop-Modell. Dabei wird, nach einer festgelegten Anzahl von neu erzeugten Datenpunkten, das Modell angepasst. Daher sollte das verwendete Modell nachtrainierbar sein oder es sollte sich schnell neu trainieren lassen. Beim AL ist zusätzlich zum Label ein Unsicherheitsmaß (oder Nützlichkeitsmaß (Settles 2009)) erforderlich, das vom Modell produziert wird. Mit dem Unsicherheitsmaß schätzt das Modell ab, wie sicher es sich

mit der Einschätzung des Labels ist. Solche Unsicherheitsinformationen können von verschiedenen Algorithmen geliefert werden, z.B. von Random-ForestKlassifikatoren oder von neuronalen Netzen die eineWahrscheinlichkeitsdichteverteilung ausgeben.

Je nachdem wie die Daten im Unternehmen vorliegen, gibt es verschiedene Arten des AL.

- Der häufigste Anwendungsfall ist das pool-basierte AL, dabei sind ungelabelte Daten in großer Zahl bereits vorhanden. (Settles 2009) Anhand der Unsicherheit des Modells wird dann ein Datenpunkt ausgewählt und gelabelt. Je nachdem wie sehr sich die Labels der unterschiedlichen Annotator:innen decken (d.h., wie hoch das InterAnnotator-Agreement ist) ist, kann es notwendig sein einen Datenpunkt von zusätzlichen Annotator:innen labeln zu lassen. Es kann aber auch je nach Anwendungsfall notwendig sein, mindestens 2 Annotator:innen einzusetzen, z.B. in medizinischen Anwendungen. Es ist vorstellbar, die vom Modell erzeugten Label (Pseudolabel) den Annotator:innen vorzuschlagen und um eine Einschätzung zu bitten. In diesem Fall sprechen wir von semi-supervised AL.
- Der zweite Fall ist das stream-basierte AL. In diesem Scenario liegen keine großen Datenbanken vor, stattdessen werden Datenpunkte in Echtzeit (online) erzeugt und es wird, ebenfalls in Echtzeit, entschieden, ob ein neues Label erzeugt werden muss. (Settles 2009)
- Ein Spezialfall des stream-basierten AL ist das query-de-novo AL, denn dabei liegen gar keine Daten vor, sondern werden synthetisch erzeugt. Dieses Vorgehen kann sinnvoll sein in Verbindung mit automatisierter Label-Generation, z.B. durch ein Roboter-gestütztes Experiment (Settles 2009). Alternativ können vom Menschen nicht nur das Label, sondern auch die Eingangsdaten geliefert werden, z.B. kann es bei Spracherkennungstrainings sinnvoll sein, dass Menschen Sätze vorlesen, die vom Computer vorgegeben werden (vgl. die Datenerhebung im Mozilla-Common-Voice Projekt). Hier bietet sich ein Gestaltungsspielraum für eine abwechslungsreichere Tätigkeit beim AL an. So könnten Menschen abwechselnd Label und Eingangsdaten erheben.

2.8 WAS IST EXPLAINABILITY UND WARUM IST DAS SINNVOLL?

Optimal ist es, wenn das Modell zusätzlich zur Unsicherheit eine Erklärung bzw. Rechtfertigung für seine Entscheidung liefern kann. Dadurch sollen Vertrauen und Identifikation mit der KI gestärkt werden, zusätzlich werden Schwächen und Grenzen des KI-Systems sichtbar und behebbar: Feedback der Annotator:innen muss sich nicht nur auf die Ausgabe des Modells beziehen, sondern kann sich auch direkt auf dessen Modi des Schlußfolgerns erstrecken (Ross, Hughes und DoshiVelez 2017). Eine Anpassbarkeit der zur Verfügung gestellten Informationen an Nutzer:innen kann ebenfalls zu einer stärkeren Zusammenarbeit mit dem KI-System führen. Es ist wichtig, dass System und Anwender:innen die gleiche Sprache sprechen und die Anwender:innen eine Intuition für das Verhalten des Systems entwickeln. Eine Form von Erklärungen können beim Labeln von Bildern z.B. Heat-Maps sein die Annotator:innen auf bestimmte Bereiche aufmerksam machen, die zur Entscheidungsfindung beigetragen haben. Siehe (Bach u. a. 2015; Selvaraju u. a. 2017). Es gibt diverse Algorithmen, die Erklärungen in einer verständlichen Darstellung liefern können;

dazu zählen Random Forests (Breiman 2001), aktivierungsbasierte Analysen oder auch backpropagation-to-input Verfahren (Mordvintsev, Olah und Tyka 2015). Einsichtige Begründungen der Schlüsse der KI sollten intuitive, z.B. natürlichsprachliche, Ausgaben sein. Je mehr Vertrauen Erklärungen geschenkt wird desto wichtiger wird die Evaluation der Angemessenheit der Vertrauenskalibrierung der Nutzer sowie die klare Kommunikation der Semantik und der Grenzen des jeweiligen Erklärverfahrens.

2.9 Wer ist verantwortlich?

Wie in (bitkom 2020; Bundesministerium Wirtschaft für und Energie 2019; Europäische Kommission 2020) erläutert, sind bei der Entwicklung und Einführung von AL-Systemen zwei Arten von Haftung zu beachten:

- Die Produkthaftung wird relevant, wenn das ALSystem als Produkt verkauft und von Endnutzern verwendet wird. Allerdings ist die Haftungssituation für nach-trainierbare KI-Systeme nicht geklärt, denn es gibt momentan keine Definition, die sagt, ob das System einen Fehler hat, wenn nach dem Nachtrainieren ein Fehlverhalten entsteht. Falls das System nur intern verwendet wird, ist die Produkthaftung nicht relevant.
- Die andere Haftungsart ist die Produzentenhaftung, die in jedem Fall relevant ist. Um dieser gerecht zu werden, müssen neben der generellen Dokumentation des Systems und seiner Risiken auch die Absicherung des Systems durch Tests dokumentiert werden. Dazu können unter anderem Softwaretests, Hardwaretests (z.B. EMV) und Betrachtungen zur maschinellen und funktionalen Sicherheit des AL-Systems gehören. Die zentrale Frage, die sich aus human-zentrierter Sicht stellt, ist dabei immer: “Was muss ich tun, damit durch mein System kein Schaden entstehen kann?” Die Antwort darauf kann heißen, durch Absicherung mit Testdatensätzen das Systemverhalten in sicherheitsrelevanten Situationen zu testen, doch momentan gibt es keine explizite Lösung für KISysteme. Es bleibt also fast immer nur, dass KISysteme und deren Output immer auch anderweitig abgesichert werden müssen, final durch ein menschliches Review.

2.10 DISKRIMINIERUNGSFREIHEIT UND FAIRNESS

Die Frage, die man sich in Bezug auf Fairness von KI-Systemen stellen sollte, ist die Frage danach, wer durch die Entscheidung der KI beinflusst wird. Hier bietet sich eine Stakeholderanalyse an, um herauszufinden, auf wen die KI direkt oder indirekt einen Einfluss hat. Diese KI-Betroffenen sollten in dem Entstehungs- und Entscheidungsprozess der KI mit einbezogen werden. Anderenfalls ergeben sich durch die Einführung eines KI-Systems ggf. Konsequenzen für Menschen, die das KI-System weder verstehen, noch einen Einfluss darauf haben. Ein solches Vorgehen steht in direktem Widerspruch zu unserer Vorstellung, wie KI human-zentriert konzipiert werden sollte, vgl. (Madiega 2019). Das gilt vor allem beim Erstellen von Datenbasen, denn durch die Auswahl der Annotator:innen und auch durch die Datenauswahl selbst, sowie durch den Bias, der den Daten eventuell inhärent ist, kann ein Bias im System entstehen, der sich später auf Menschen auswirkten kann, siehe (Van Houten

2020).

Beim AL könnten durch Inter-Annotator-Kommunikation Echokammern entstehen, die einen Bias verstärken könnten, dazu ist mehr Forschung notwendig. Viele dieser Aspekte finden sich bereits in aktueller Literatur (Bundesministerium für Bildung und Forschung, Bundesministerium für Wirtschaft und Energie und Bundesministerium für Arbeit und Soziales 2018; Burt 2020; Orwat und Antidiskriminierungsstelle des Bundes 2019), es sind aber auch hier weitere Untersuchungen erforderlich. Was kann und geschehen sollte, um spezifisch bei AL das Entstehen von Bias zu vermeiden, ist ein aktuelles Forschungsfeld

LITERATURVERZEICHNIS

Bach, Sebastian u. a. (2015). “On pixel-wise explanations for non-linear classifier decisions by layerwise relevance propagation”. In: *PloS one* 10.7, e0130140.

Bakker, Arnold B. (2005). “Flow among music teachers and their students: The crossover of peak experiences”. In: *Journal of vocational behavior* 66.1, S. 26–44. (2008). “The work-related flow inventory: Construction and initial validation of the WOLF”. In: *Journal of vocational behavior* 72.3, S. 400– 414.

Bassi, Marta und Antonella Delle Fave (2012). “Optimal experience among teachers: New insights into the work paradox”. In: *The Journal of psychology* 146.5, S. 533–557.

Berka, Chris u. a. (2007). “EEG correlates of task engagement and mental workload in vigilance, learning, and memory tasks”. In: *Aviation, space, and environmental medicine* 78.5, B231–B244.

Bessiere, Katie u. a. (2004). “Social and psychological influences on computer user frustration”. In: *Media access: Social and psychological dimensions of new technology use,* S. 91–103.

bitkom (2020). *Stellungnahme: Rechtsfragen der digitalisierten Wirtschaft: Haftung für Systeme Künstlicher Intelligenz.*

Breiman, Leo (2001). “Random forests”. In*: Machine learning* 45.1, S. 5–32.

Bundesministerium für Bildung und Forschung, Bundesministerium für Wirtschaft und Energie und Bundesministerium für Arbeit und Soziales (2018). *Strategie Künstliche Intelligenz der Bundesregierung.*

Bundesministerium Wirtschaft für und Energie (2019). *Künstliche Intelligenz und Recht im Kontext von Industrie 4.0.*

Burt, Andrew (2020). “How to Fight Discrimination in AI”. In: *Harvard Business Review*

Chakraborty, Shayok, Vineeth Balasubramanian und Sethuraman Panchanathan (2015). “Adaptive Batch Mode Active Learning”. In: *IEEE Transactions on Neural Networks and Learning Systems* 26.8, S. 1747–1760. doi: 10.1109/TNNLS. 2014.2356470.

Csikszentmihalyi, Mihaly (1975). *Beyond boredom and anxiety. Jossey-Bass.*

Csikszentmihalyi, Mihaly und Judith LeFevre (1989). “Optimal experience in work and leisure.” In*: Journal of personality and social psychology* 56.5, S. 815.

Emanuel, Federica, Lara Colombo und Margherita Zito (2016). “Flow at work in Italian journalists: differences between permanent and freelance journalists”. In: *Flow at work in Italian journalists: differences between permanent and freelance journalists,* S. 26–46.

Europäische Kommission (2020). *Bericht über die Auswirkungen künstlicher Intelligenz, des Internets der Dinge und der Robotik in Hinblick auf Sicherheit und Haftung*.

Fagerlind, Anna-Carin u. a. (2013). “Experience of work-related flow: does high decision latitude enhance benefits gained from job resources?” In: *Journal of vocational behavior* 83.2, S. 161–170.

Fleischer, Nancy L. u. a. (2013). “Public health impact of heat-related illness among migrant farmworkers”. In: *American journal of preventive medicine* 44.3, S. 199–206.

Hüttges, A., A. Müller und P. Richter (2005). “Gesundheitsförderliche Arbeitsgestaltung durch Kurzpausensysteme: Ein Ansatz an der Schnittstelle von Verhaltens-und Verhältnisprävention”. In: *Wirtschaftspsychologie* 7.3, S. 36–43.

Kim, Sooyeol, YoungAh Park und Lucille Headrick (2018). “Daily micro-breaks and job performance: General work engagement as a cross-level moderator.” In: *Journal of Applied Psychology* 103.7, S. 772.

Madiega, Tambiama (2019). *EU guidelines on ethics in artificial intelligence: Context and implementation.*

Maeran, Roberta und Francesco Cangiano (2013). “Flow experience and job characteristics: Analyzing the role of flow in job satisfaction”. In: *TPMTesting, Psychometrics, Methodology in Applied Psychology* 20.1, S. 13–26.

Mordvintsev, Alexander, Christopher Olah und Mike Tyka (2015*). Inceptionism: Going Deeper into Neural Networks.*

Orwat, Carsten und Antidiskriminierungsstelle des Bundes (2019). *Diskriminierungsrisiken durch Verwendung von Algorithmen.*

Peifer, Corinna und Gina Wolters (2017). “Bei der Arbeit im Fluss sein: Konsequenzen und Voraussetzungen von Flow-Erleben am Arbeitsplatz”. In: *Editorial: Positive Psychologie im Kontext von Arbeit und Organisation* 19.3, S. 6–22.

Richter, Peter und Winfried Hacker (2000). *Belastung und Beanspruchung: Stress, Ermüdung und Burnout im Arbeitsleben.* Asanger.

Ross, Andrew Slavin, Michael C. Hughes und Finale Doshi-Velez (2017). “Right for the right reasons: Training differentiable models by constraining their explanations”. In: *arXiv preprint arXiv:1703.03717*.

Selvaraju, Ramprasaath R. u. a. (2017). “Gradcam: Visual explanations from deep networks via gradient-based localization”. In: *Proceedings of the IEEE international conference on computer vision,* S. 618–626.

Settles, Burr (2009). *Active Learning Literature Survey*. Computer Sciences Technical Report 1648. University of Wisconsin–Madison.

Singh, Ankita und Shayok Chakraborty (2020). “Deep Active Transfer Learning for Image Recognition”. *In: 2020 International Joint Conference on Neural Networks (IJCNN),* S. 1–9. doi: 10.1109/IJCNN48605.2020.9207391.

Steiger, Miriah u. a. (2021). “*The psychological well-being of content moderators”. In: Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems, CHI.* Bd. 21.

Tausch, Alina und Corinna Peifer (2019). “Auswirkungen von Autonomie auf Flow, Motivation und Leistung: Eine Studie im Schaltanlagenbau”. In: *Wirtschaftspsychologie* 21.4, S. 83–100.

Van Houten, Henk (2020). *For fair and equal healthcare, we need fair and bias-free AI*.

Verordnung (EU) 2016/679 (2016). *Verordnung (EU) 2016/679 des Europäischen Parlaments und des Rates vom 27. April 2016 zum Schutz natürlicher Personen bei der Verarbeitung personenbezogener Daten, zum freien Datenverkehr und zur Aufhebung der Richtlinie 95/46/EG (Datenschutz-Grundverordnung) (Text von Bedeutung für den EWR).*

Warm, Joel S., Gerald Matthews und Victor S. Finomore Jr (2008). “Vigilance, workload, and stress”. In: *Performance under stress*. CRC Press.