Lossos, Christian; Geschwill, Simon; Morelli, Frank

Offenheit durch XAI bei ML-unterstützten Entscheidungen: Ein Baustein zur Optimierung von Entscheidungen im Unternehmen?

HMD Praxis der Wirtschaftsinformatik

Provided in Cooperation with:
Springer Nature

Suggested Citation: Lossos, Christian; Geschwill, Simon; Morelli, Frank (2021) : Offenheit durch XAI bei ML-unterstützten Entscheidungen: Ein Baustein zur Optimierung von Entscheidungen im Unternehmen?, HMD Praxis der Wirtschaftsinformatik, ISSN 2198-2775, Springer Fachmedien Wiesbaden, Wiesbaden, Vol. 58, Iss. 2, pp. 303-320, https://doi.org/10.1365/s40702-021-00707-1

HMD (2021) 58:303–320
https://doi.org/10.1365/s40702-021-00707-1

SCHWERPUNKT

Offenheit durch XAI bei ML-unterstützten Entscheidungen: Ein Baustein zur Optimierung von Entscheidungen im Unternehmen?

Christian Lossos · Simon Geschwill · Frank Morelli

Eingegangen: 18. Oktober 2020 / Angenommen: 31. Januar 2021 / Online publiziert: 3. März 2021

Zusammenfassung Künstliche Intelligenz (KI) und Machine Learning (ML) gelten gegenwärtig als probate Mittel, um betriebswirtschaftliche Entscheidungen durch mathematische Modelle zu optimieren. Allerdings werden die Technologien häufig in Form von „Black Box"-Ansätze mit entsprechenden Risiken realisiert. Der Einsatz von Offenheit kann in diesem Kontext mehr Objektivität schaffen und als Treiber für innovative Lösungen fungieren. Rationale Entscheidungen im Unternehmen dienen im Sinne einer Mittel-Zweck-Beziehung dazu, Wettbewerbsvorteile zu erlangen. Im Sinne von Governance und Compliance sind dabei regulatorische Rahmenwerke wie COBIT 2019 und gesetzliche Grundlagen wie die Datenschutz-Grundverordnung (DSGVO) zu berücksichtigen, die ihrerseits ein Mindestmaß an Transparenz einfordern. Ferner sind auch Fairnessaspekte, die durch Bias-Effekte bei ML-Systemen beeinträchtigt werden können, zu berücksichtigen. In Teilaspekten, wie z. B. bei der Modellerstellung, wird in den Bereichen der KI und des ML das Konzept der Offenheit bereits praktiziert. Das Konzept der erklärbaren KI („Explainable Artificial Intelligence" – XAI) vermag es aber, das zugehörige Potenzial erheblich steigern. Hierzu stehen verschiedene generische Ansätze (Ante hoc-, Design- und Post-hoc-Konzepte) sowie die Möglichkeit, diese untereinander zu kombinieren, zur Verfü-

C. Lossos
Stuttgart, Deutschland
E-Mail: christian.lossos@gmail.com

S. Geschwill
Hockenheim, Deutschland
E-Mail: Simon.Geschwill@web.de

F. Morelli (✉)
Hochschule Pforzheim – Gestaltung, Technik, Wirtschaft und Recht,
Tiefenbronnerstr. 65, 75175 Pforzheim, Deutschland
E-Mail: frank.morelli@hs-pforzheim.de

gung. Entsprechend müssen Chancen und Grenzen von XAI systematisch reflektiert werden. Ein geeignetes, XAI-basiertes Modell für das Fällen von Entscheidungen im Unternehmen lässt sich mit Hilfe von Heuristiken näher charakterisieren.

Schlüsselwörter Betriebswirtschaftliche Entscheidungen · Erklärbare Künstliche Intelligenz (XAI) · Maschinelles Lernen (ML) · Heuristik · Offenheit

Openness Through XAI in ML-Assisted Decisions: A Building Block for Optimizing Enterprise Decision-Making?

Abstract Artificial Intelligence (AI) and Machine Learning (ML) are currently considered to be effective tools to optimize business decisions by applying mathematical models. However, they are often implemented as "black box" approaches with corresponding risks. In this context, the usage of openness can create more objectivity and act as a driver for innovative solutions. Rational decisions within the company serve the purpose of gaining competitive advantages in the sense of a means-ends relationship. In terms of governance and compliance, regulatory frameworks like COBIT 2019 and legal foundations such as the General Data Protection Regulation (GDPR) must be taken into account, which require a minimum level of transparency. Furthermore, fairness aspects, which can be affected by bias effects in ML models, have also to be considered. In some aspects, such as in model development, openness is already practiced in the areas of AI and ML. However, the concept of Explainable Artificial Intelligence (XAI) is able to significantly increase potentials. Various generic approaches (ante hoc, design and post-hoc concepts) are available for this purpose, as well as the possibility of combining them with each other. Accordingly, the opportunities and limitations of XAI must be systematically reflected upon. An appropriate XAI-based model for decision making in companies can be characterized by support of heuristics.

Keywords Business decisions · Explainable Artificial Intelligence (XAI) · Machine Learning (ML) · Heuristics · Openness

1 Offenheit und maschinelles Lernen (ML) für betriebswirtschaftliche Entscheidungen

Das Konzept der Offenheit lässt sich generell durch Transparenz, Zugang, Partizipation und Demokratie charakterisieren. Die Offenlegung von Informationen ermöglicht ein höheres Maß an Objektivität durch intersubjektive Überprüfbarkeit. Für die Wirtschaftsinformatik kann Offenheit als Treiber für innovative Lösungen fungieren. Ferner besteht die Möglichkeit, dass Offenheit das Ergebnis innovativer IT-Ansätze ist. (Schlagwein et al. 2017, S. 297).

Für betriebswirtschaftliche Entscheidungen stellt sich grundsätzlich die Frage nach dem Nutzen von Offenheit und IT-Unterstützung. Unter Berücksichtigung der begrenzten kognitiven Fähigkeiten des Menschen sind optimale Entscheidungen im betriebswirtschaftlichen Kontext i. d. R. nicht zu bewältigen. Durch den Einsatz von

Künstlicher Intelligenz (KI) und dabei insbesondere von maschinellem Lernen („Machine Learning", ML) im Unternehmen verspricht man sich mehr Effektivität und Effizienz. Offenheit lässt sich mit dem Treffen rationaler Entscheidungen verknüpfen. Diese weisen dann eine eindeutige und nachvollziehbare Mittel-Zweck-Relation auf: Aus mehreren Handlungsalternativen ist diejenige auszuwählen, welche der bestmöglichen Umsetzung unternehmerischer Zielsetzungen dient.

Unter KI wird nachfolgend die Fähigkeit zur Simulation von intelligentem Verhalten durch Systeme verstanden, indem Umgebungsdaten analysiert und daraus resultierende Aktionen zur Erreichung definierter Ziele in der Umgebung abgeleitet werden. Entsprechende Anwendungsfälle existieren sowohl in Software- (z. B. Spracherkennung) als auch in Hardware-Anwendungen (z. B. autonomes Fahren). (HLEG-AI 2021, S. 1; Merriam-Webster 2020; Russell und Norvig 2016, S. viii) Gegenwärtig diskutiert man im KI-Kontext in besonderem Maße Verfahren des ML, die im weiteren Verlauf fokussiert werden. Dabei handelt es sich um Algorithmen, die eigenständig und kontinuierlich ihre Leistungen durch Lernen verbessern. Dies wird realisiert, indem Muster in verfügbaren Datensätzen identifiziert, das dadurch erhaltene Wissen in einem statischen Modell berücksichtigt und dies auf neue Daten angewendet wird. Systeme entwickeln dadurch neuartige, subjektiv anmutende Lösungswege unter Berücksichtigung ihrer Umgebung. (Goodfellow et al. 2016, S. 2 f.; Merriam-Webster 2021; European Commission 2018, S. 11).

Beim ML-Ansatz soll eine datenbasierte Fundierung die Realität abbilden und optimale betriebswirtschaftliche Entscheidungen gewährleisten. Dabei stellt sich jedoch die Frage nach der Realistik dieses Vorhabens: Zum einen ist durch den Einsatz von ML die Realisierung transparenter und erklärbarer maschineller Entscheidungen nicht unmittelbar gewährleistet. Bei ML handelt es sich vielmehr oftmals um „Black-Box"-Ansätze. Diese Verfahren legen keine Korrelationen bzw. kausalen Zusammenhänge zwischen Variablen offen. Dadurch bleiben mögliche Bias-Effekte unentdeckt. Kombiniert man den ML-Ansatz mit dem Offenheitsprinzip, erscheint ein höheres Maß an Objektivität gewährleistet. Aus diesem Grund hat sich in jüngerer Vergangenheit die Erklärbarkeit von künstlicher Intelligenz („Explainable Artificial Intelligence", XAI) als Forschungszweig etabliert (Arrieta et al. 2020). Die Ziele von XAI bestehen in der Erstellung transparenter und erklärbarer maschineller Entscheidungen sowie in der verständlichen Gestaltung der Funktions- und Arbeitsweisen von KI-Systemen. (Lipton 2016) Zum anderen sind im Rahmen der betriebswirtschaftlichen Entscheidungsfindung auf Basis von großen Datenmengen mit multiplen Korrelationen oftmals personenbezogene Daten zu berücksichtigen. Entsprechend müssen Unternehmen externe Regularien wie z. B. COBIT 2019 oder die Datenschutz-Grundverordnung (DSGVO) berücksichtigen und im Prozess der Datenbereitstellung und -auswertung ein gewisses Maß an Transparenz gewährleisten. Der vorliegende Artikel zeigt Möglichkeiten und Grenzen der Kombination von Offenheit und ML auf und gibt Gestaltungsvorschläge in Form von heuristischen Empfehlungen.

2 Entscheidungen im Unternehmen

Ein zu Entscheidungen gehörendes Zielsystem ist bei der Verfolgung mehrerer, zum Teil widersprüchlicher Ziele durch Konflikte geprägt. Der Entscheidungsbegriff umfasst i. w. S. nicht nur einen Entschluss, sondern repräsentiert einen sich im Zeitablauf vollziehenden Prozess (Problemformulierung, Präzisierung des Zielsystems, Erforschen des Handlungsspektrums, Alternativenauswahl sowie Entscheidungen in der Realisierungsphase). (Laux et al. 2018, S. 12) Dabei sind Interdependenzen zwischen diesen Aktivitäten in einem holistischen Verständnis für die Optimierung zu berücksichtigen. Im Sinne der Performanz ist es wichtig, dass die Entscheidungsträger Informationen möglichst objektiv verarbeiten um Fehlentscheidungen zu vermeiden.

Verschiedene Wissenschaftsdisziplinen beschäftigen sich mit der Formulierung und Lösung von Entscheidungsproblemen. Die Entscheidungstheorie hat sich dabei als interdisziplinäres Forschungsgebiet herausgebildet, das sich mit dem Entscheidungsverhalten von Individuen und Gruppen (Entscheidungsgremien) befasst. Das menschliche Gehirn decodiert physikalische Größen, interpretiert sie und konstruiert daraus subjektiv eine zusammenhängende Realität. Große und komplexe Datenmengen können die kognitiven Fähigkeiten übersteigen und die daraus resultierende Informationsüberlastung führt nicht zu besseren Entscheidungen im Vergleich zu Methoden der rationalen Entscheidungsfindung. (Chlupsa 2017 und die dort zitierte Literatur, S. 3) Menschliche Entscheidungen werden immer sowohl durch explizite und implizite Emotionen bzw. Gefühle als auch Motive beeinflusst. Motive befähigen Menschen auf ähnliche Situationen mit gleichen Mustern zu interagieren. Basierend auf Wissen und Erfahrung eröffnet die Intuition von Fachexperten i.S.e. impliziten Motivation durch ganzheitliche Assoziationen ein Potenzial für nachhaltiges und verantwortliches Handeln. Um einen Expertenstatus zu erreichen, ist ein direktes Feedback bedeutsam. (Chlupsa 2017, S. 31) Bei Gruppenentscheidungen werden als zusätzliche Kriterien Kommunikation und Abstimmung zwischen den Beteiligten berücksichtigt. Typischerweise differenziert man in diesem Zusammenhang den Informations- und den Abstimmungsprozess. Es gibt jedoch auch kritische Stimmen zur Effizienz von Gruppenentscheidungen. (Chlupsa 2017, S. 274) Forschungsarbeiten über Gruppenentscheidungen greifen das Problem eines „gerechten“ Interessensausgleichs bzw. die Frage der Fairness von Präferenzordnungen auf.

Im Hinblick auf eine wirksame Kombination von menschlichen Entscheidungsträgern mit ML-Algorithmen spielt eine Vielzahl von Kriterien eine wichtige Rolle: Geschäftsvorfall und Wertschöpfung, Datensammlung und Datenquellen, Datenintegration und Datenqualität, ML-Methoden, eingesetzte Software und die Modellqualität, Kompetenzen der involvierten Personen und der Stakeholder sowie Kosten-Nutzenüberlegungen. (Neifer et al. 2021, S. 5400 ff.) Eingesetzte Black-Box-Systeme lassen definitionsgemäß keine Transparenz über innere Verhaltensweisen zu. Umgekehrt gewährleisten White-Box-Systeme per se jedoch keine Nachvollziehbarkeit der Entscheidungen durch menschliche Entscheidungsträger. Dies hängt zum einen von den methodischen Kenntnissen und Fähigkeiten der Entscheidungsträger als auch der Stakeholder ab. Mentale Modelle, d. h. wie Menschen KI bzw. ML

wahrnehmen, spielen zum anderen eine wichtige Rolle. Beispielsweise ist davon auszugehen, dass Menschen nach Erklärungen suchen, wenn ein System zu unerwarteten Resultaten gelangt. Ein adäquates konnotatives Verständnis schafft hinreichendes Vertrauen und Akzeptanz sowie Bewusstsein für das Ergebnis. (Alizadeh et al. 2020).

Rationale unternehmerische Entscheidungen sind Mittel zur Zielerreichung (Schencking 2018, S. 120 ff.), i. d. R. um Wettbewerbsvorteile zu erzielen. Sie setzen mehrere potenzielle Handlungsalternativen voraus und sollen die bestmögliche Umsetzung unternehmerischer Zielsetzungen gewährleisten. (Holtmann 2008, S. 20 f.) In Unternehmen ist zwischen strategischen und operativen Entscheidungen zu differenzieren: Strategische Entscheidungen sind immer zukunftsbezogen, mit entsprechenden Unsicherheiten und Prognoseproblemen behaftet und insbesondere auf Effektivität ausgerichtet. (Amann und Petzold 2014, S. 49 f.) Operative Entscheidungen erweisen sich demgegenüber als überwiegend gegenwartsbezogen und sind auf Effizienz ausgerichtet. (Amann und Petzold 2014, S. 153 f.) Sie eignen sich tendenziell besser für den Einsatz von ML-Verfahren: Normative Aspekte wie Werte, Überzeugungen und grundlegende Einstellungen spielen für strategische Überlegungen eine wichtige Rolle. Die hieraus abgeleiteten Ziele existieren oftmals nur „verdeckt“. Operative Entscheidungen finden vergleichsweise in einem klaren Kontext von kostenmäßigen, zeitlichen und/oder qualitativen Restriktionen statt. Auch sind die Vollständigkeit der Entscheidungsalternativen und die Auswirkung auf Folgeentscheidungen für operative Sachverhalte leichter zu beurteilen. Einen weiteren Aspekt beinhaltet die Menge an verfügbaren Daten, die für eine Fokussierung auf operative Problemstellungen spricht.

Unternehmerische Entscheidungen werden durch normative Rahmenbedingungen reglementiert. Die dauerhafte Einhaltung rechtlicher und faktischer Ordnungsrahmen stellt dabei den Fokus für Governance und Compliance dar. Eine Adressierung von Offenheit erfolgt durch verschiedene Standards und gesetzliche Vorgaben, die in einer dezidierten Analyse im Kontext ML-gestützter Entscheidungen untersucht werden müssen.

3 Regulatorische Rahmenaspekte

IT-Governance verfolgt das Ziel, die zunehmend komplexer werdende Unternehmens-IT aus ganzheitlicher Perspektive zu steuern: Es sollen sowohl ein Wertbeitrag erbracht als auch Risikoaspekte berücksichtigt werden. Diese Prämisse ist für zugehörige Unternehmensentscheidungen relevant.

COBIT 2019, ein international anerkanntes Framework zur IT-Governance, forciert Offenheit an verschiedenen Stellen: Im Kontext seines aus fünf Modulen bestehenden Kern-Modells zeigt sich die Forderung nach Offenheit vor allem in den Modulen „Evaluate, Direct and Monitor (EDM)“, „Deliver, Service and Support (DSS)“ sowie „Monitor, Evaluate and Assess (MEA)“. (COBIT 2019) Während die Risiko-Optimierung im Rahmen von EDM eine implizite Offenheit im Rahmen des artikulierten Risikoverständnisses darstellt, beinhaltet das Governance-Ziel der Stakeholder-Einbindung bei IT-Performanz-Messungen im Rahmen der Trans-

parenz-Forderung eine explizite Anforderung an Offenheit. Eine ML-gestützte Performanz-Messung sollte daher ein hohes Maß an Transparenz beinhalten. Das DSS-Modul erfordert u. a. eine Kontrolle der Geschäftsprozesse und lässt sich im Kontext von ML-Entscheidungen auf die Kontrolle von Eingabe, Durchsatz und Ausgabe anwenden. Dabei kann Offenheit helfen diese Anforderungen zu kontrollieren und sicherzustellen. Auch die Identifizierung von Problemen und deren Ursache lässt sich als implizite Forderung nach Offenheit interpretieren. Eine zusätzliche Forderung nach Offenheit im Rahmen von Performanz-Monitoring findet sich, neben der Forderung nach einem Konformitäts-Monitoring externer und interner Anforderungen, ebenfalls im Modul MEA: Prozesse sollen mit Hilfe von festgelegten Performanz- und Konformitäts-Metriken evaluiert werden. Bei ML-gestützten Entscheidungen ergibt sich hieraus die Anforderung zur Transparenzschaffung gegenüber einer potenziellen „Black-Box"-Entscheidung.

Die rechtlichen Rahmenbedingungen in Bezug auf KI befinden sich aktuell noch im Entwicklungsstadium, allerdings lässt sich eine Tendenz zur Forderung von Offenheit auf europäischer Ebene insbesondere durch die „Assessment List for Trustworthy AI" (Alliance 2020) erkennen. Das Erreichen von vertrauenswürdiger KI soll durch Nachvollziehbarkeit, Erklärbarkeit und Kommunikation mit Benutzern erreicht werden. Offenheit manifestiert sich dabei im Aspekt der Erklärbarkeit, wobei nicht nur die Entscheidung dem User aufgezeigt, sondern auch kontinuierlich das Verständnis der KI-Entscheidung durch die Nutzer analysiert werden soll. Es muss dabei auch sichergestellt werden, dass Benutzer mit den Risiken, Limitierungen und Ziel des Systems vertraut sind. (Alliance 2020).

Im Rahmen der DSGVO findet sich die Verwendung von Offenheit und Transparenz explizit in Artikel 5 Absatz 1 (a). Die verwendete Definition von Transparenz lässt sich mit Hilfe von Erwägungsgrund 58 näher charakterisieren: Der zugehörige Fokus liegt auf Zugang, Verständlichkeit und Prägnanz (Sartor und Lagioia 2020). In diesem Kontext ist allerdings ein Unterschied zwischen transparenter und erklärbarer KI zu erkennen: Im Rahmen von ML-Systemen kann eine Unterteilung in Transparenz nach dem Zeitpunkt der Informationsverarbeitung vorgenommen worden. Die Artikel 13 Absatz 2 (f) und 14 Absatz 2 (g) DSGVO beziehen sich auf die Verarbeitung von Daten im Vorfeld und fordern eine Informationspflicht über die Existenz von Systemen mit automatisierter Entscheidungsfindung sowie über die Logik des Systems und vorgesehene Konsequenzen. Idealerweise soll hierbei der Nutzer bereits über die Daten, welche das System verarbeitet, sowie über deren Priorisierung informiert werden. Zudem sind der durch das System berechnete Zielwert sowie abgeleitete Konsequenzen an den Nutzer zu kommunizieren. Artikel 15 Absatz 1 DSGVO nimmt Bezug auf bereits verarbeitete Daten und entspricht den oben genannten Anforderungen. Detaillierte Ausführungen, ob Betroffene nur mit generellen Informationen versorgt werden müssen oder ob es individueller Erklärungen bedarf, existieren allerdings nicht.

4 Künstliche Intelligenz (KI) und maschinelles Lernen (ML) im Kontext von Offenheit

KI und ML finden weltweit in vielen Branchen und Disziplinen Anwendung. Der gesellschaftliche und wirtschaftliche Alltag wird bereits heute durch unterschiedliche Ausprägungen, wie z. B. Spracherkennung bei digitalen Assistenten, durch KI geprägt. Die Komplexität der Anwendungsfälle kann dabei stark variieren und in einzelnen Fällen bessere Ergebnisse als menschliche Fachexperten erzielen wie beispielsweise in einer medizinischen Studie zur Erkennung von Hautkrebs nachgewiesen werden konnte. (Haenssle et al. 2018) Analysten der Unternehmensberatung McKinsey & Company bewerten die Auswirkungen von KI auf die Weltwirtschaft höher als die durch die Erfindung der Dampfmaschine im 19. Jahrhundert und dem Aufkommen von Informations- und Kommunikationstechnologien im 21. Jahrhundert. (Bughin et al. 2018) Häufig berücksichtigt man dabei nicht, dass KI lediglich als Überbegriff für eine eigenständige Disziplin innerhalb der Informatik fungiert, welche aus diversen Teilgebieten besteht. So werden bereits seit vielen Jahren Experten- bzw. regelbasierte Systeme kommerziell eingesetzt, welche auf klassischen Paradigmen der Informatik in Form von Wenn-Dann-Regeln beruhen.

Die Methodik zur Erstellung von ML-basierten Modellen unterscheidet sich jedoch stark von der Vorgehensweise bei der klassischen Softwareentwicklung, da ein großer Fokus auf den zugrunde liegenden Daten des jeweiligen Anwendungsfalls liegt. Mit CRISP-DM (Cross-Industry Standard Process for Data Mining) wurde im Jahr 2000 ein offener Entwicklungsansatz für Data Mining entwickelt, der sich bis heute als Industriestandard im Bereich Data Science etabliert hat und in sechs Phasen gegliedert ist. Zu Beginn der Entwicklungsaktivitäten wird der betriebswirtschaftliche Kontext sowie das Themenfeld (z. B. Ziele bzw. Problemstellungen, Abhängigkeiten) des jeweiligen Anwendungsfalls analysiert („Business Understanding“). Die Fokussierung auf die zur Verfügung stehenden Daten während der Entwicklung spiegelt sich in den folgenden Abschnitten „Data Understanding“ sowie „Data Preparation“: Zunächst unterzieht man die Daten einer detaillierten Analyse (u. a. Formate, Struktur-Redundanzen, Abhängigkeiten) und anschließend bereitet man das Resultat für die weitere Verarbeitung vor (u. a. Formatierung, Zusammenfassung). In Bezug auf die Verteilung von Ressourcen während des Entwicklungsprozesses entfällt ein Großteil der Aktivitäten auf dieses Segment. Anschließend werden auf Basis der bearbeiteten Daten verschiedene Modelle erzeugt („Modeling“), deren Performance anschließend evaluiert („Evaluation“) und ein finales Modell in den Anwendungskontext implementiert („Deployment“). Die Abschnitte von CRISP-DM sind jedoch nicht als sequenzielle Abfolge konzipiert. Vielmehr erlauben sie Freiheiten für Rücksprünge zwischen den verschiedenen Phasen. (Chapman et al. 2000) In der Zwischenzeit sind weitere Vorgehensweisen zur Entwicklung von ML-Modellen veröffentlicht worden, darunter auch CPMAI (Cognitive Project Management for Artificial Intelligence). Dieses Konzept ergänzt CRISP-DM um verschiedene Attribute, welche die einzelnen Aktivitäten detaillierter beschreiben.

Die Umsetzung des Entwicklungsansatzes lässt sich in Abhängigkeit von situativen Gegebenheiten (z. B. Anwendungsfall und Organisation) variieren, beispielsweise durch die Anreicherung mit agilen Methoden und Softwarekomponenten wie

z. B. ML-Bibliotheken. Insbesondere in Bezug auf die verwendete Software steht Entwicklern in der Praxis eine große Anzahl an umfangreichen und offenen Frameworks zur Verfügung (z. B. Keras, TensorFlow und Scikit-Learn). Zudem besteht eine große Internet-Community für den Bereich ML, welche Hilfestellungen und Unterstützung bei der Umsetzung von Projekten bietet.

Unternehmen veröffentlichen entgegen dem Offenheitsprinzip häufig jedoch keine Details zur Umsetzung von KI- bzw. ML-Projekten und begründen dies mit Vertraulichkeitsaspekten. Betroffen davon sind i. d. R., neben den verwendeten Softwarekomponenten, die Spezifikation der Datengrundlage bzw. der zugehörigen Qualitätsstandards, die für den Erfolg des späteren ML-Modells ausschlaggebend sind. Als Konsequenz ergibt sich hieraus z. B. im B2C-Bereich, dass es für den Endkunden oftmals nicht ersichtlich ist, wie der Umgang mit personenbezogenen Daten durch automatisierte Systeme im jeweiligen Unternehmen erfolgt. In einzelnen Branchen und Fällen besteht allerdings ein Anrecht der Betroffenen auf Transparenz und Offenheit. Neben den bereits definierten regulatorischen Anforderungen definieren verschiedene Unternehmen interne Standards bei der Entwicklung und dem Betrieb von ML-Systemen. Dies ist beispielsweise auch unter dem Kontext zu sehen, dass der Bezug von ML-Software von Dritten zu potenziellen Risiken führt, da angewendete Vorgehensweisen im Rahmen der Entwicklung und verwendete Datensätze nicht transparent sind.

Die Prüfung auf Wirksamkeit von eingesetzten ML-Modelle kann auf Basis verschiedener, öffentlicher Aspekte erfolgen. Neben den eingangs beschriebenen internationalen und nationalen gesetzlichen Rahmenbedingungen existieren öffentliche Kompendien, welche die Prüfung von KI-Systemen anwendungsbezogen und unter Einbindung geltender gesetzlicher Vorgaben beschreiben (Lossos et al. 2019). Neben vollständig öffentlichen Ressourcen zur Prüfung von KI-Systemen lassen sich weitere Informationen über spezifische Fach- und Berufsverbände beziehen. Beispielsweise stellt der internationale Berufsverband Information Systems Audit and Control Association (ISACA) für Mitglieder Dokumente zur Verfügung, welche die Prüfung von KI- und ML-Systemen unterstützen. (Clark 2018) Wie auch bei der Entwicklung und dem Betrieb von ML-Modellen definieren individuelle Unternehmen darüber hinaus eigene Vorgehensweisen zur Prüfung der Wirksamkeit von ML-Modellen, welche jedoch i. d. R. nicht veröffentlicht werden. Diese basieren im Grundsatz auf einer Kombination von technischen und wirtschaftlichen Aspekten in Analogie zu den bereits beschriebenen Vorgehensweisen wie CRISP-DM und CPMAI und reichern diese mit weiteren Spezifika an. Hierunter fallen z. B. die Prüfung von KI-Projekten und -Prozessen auf Unternehmens- oder Bereichsebene.

Informationen zum operativen Einsatz von ML werden durch Unternehmen häufig im Rahmen von Use Case-Szenarien vorgestellt, die keine konkreten Details zu eingesetzten Software-Komponenten oder Entwicklungsaktivitäten beinhalten. Diese Vorgehensweise wird oftmals mit der Sicherung von Wettbewerbsvorteilen begründet. Im Kontext von ML-Projekten basieren diese jedoch insbesondere auf der jeweils zur Verfügung stehenden bzw. verwendeten Datengrundlage. Entsprechend sollten i. d. R. keine Einschränkungen bzgl. der Veröffentlichung eingesetzter Softwarekomponenten bestehen, da man im ML-Umfeld meist auf Open-Source-Tools zurückgreift. Die Datengrundlage hingegen bildet ein enormes Risikopotenzi-

al durch fehlerhafte oder unvollständige Informationen, die ein verzerrtes Bild der Realität darstellen können. Hieraus resultieren ggf. fehlerhafte oder diskriminierende Entscheidungen, die einzelne Personengruppen benachteiligen. Entsprechende Vorfälle haben in der gesellschaftlichen Diskussion teilweise ein negatives Bild von KI-Anwendungen geprägt: So hatte bei Amazon fehlende Transparenz im Jahre 2015 dazu geführt, dass ein Recruiting-Tool wieder abgeschafft werden musste, da es die Lebensläufe von Männern automatisiert besser bewertete als von Frauen. (Dastin 2018) Dies begründet u. a. den Sachverhalt, dass viele Forschungs- und Entwicklungsansätze im Bereich Data Science ihren Fokus auf die Datenbasis legen, um entsprechende Risiken zu minimeren.

Die Nutzung von Offenheit im Rahmen der Entwicklung und dem Betrieb von ML-Algorithmen vermag Hinweise auf potenzielle Verzerrungen geben. In Datensätzen können vielfältige Bias-Formen entstehen (Mehrabi et al. 2019; Verma und Rubin 2018) und sich darüber hinaus gegenseitig beeinflussen. Ein Algorithmus, der auf einem Datensatz mit Bias entwickelt wurde, kann beispielsweise zu einer Verstärkung der Verzerrung führen, indem die ML-basierten Entscheidungen die Entstehung der zukünftigen Daten beeinflussen, welche man für zukünftige Trainingszyklen des Modells benutzt. Weitere Formen des Bias, wie Popularitätsbias, Rangbias oder entwickeltem Bias entstehen, indem der Algorithmus die Interaktionsmöglichkeiten des Benutzers beeinflusst. So führt bspw. der Rangbias in Suchmaschinen dazu, dass häufig angeklickte Seiten weiter oben in den Resultaten landen und dadurch auch häufiger als vermeintlich gewünschtes Ergebnis vorgeschlagen werden, was ihren Rang als relevantes Resultat noch weiter verstärkt. Benutzer ihrerseits vermögen es wiederum die zugrundeliegenden Daten für den Algorithmus durch Bias-Formen wie verlinktem, Inhalte-erstellendem oder Verhaltens-Bias zu beeinflussen. Datensätze mit zeitlichem Bias, Aggregationsbias oder historischem Bias beeinflussen den Algorithmus in seiner Vorhersage und Genauigkeit. Speziell ein historischer Bias manifestiert bereits bestehende Ungleichverteilungen, indem bspw. Suchresultate einen ungleichverteilten Datensatz wie das Verhältnis von weiblichen und männlichen Führungskräften liefern und leicht zu falschen Schlussfolgerungen bei menschlichen Entscheidungsträgern führen.

Verstärkte Bestrebungen zur Förderung von Offenheit in verschiedenen Bereichen der ML können Chancen für Unternehmen generieren und die technologische Entwicklung beschleunigen: Bereits bei der Entwicklung von ML-Algorithmen lassen sich durch transparente Kommunikation Ungleichverteilungen in Daten und Modellvorhersagen leichter identifizieren. Eine Möglichkeit, Offenheit im Kontext des operativen Einsatzes von ML zu fördern und gleichzeitig die Vertraulichkeit bzgl. der verwendeten Datengrundlage zu gewährleisten, besteht im Aufbau von unternehmensübergreifenden Partnerschaften. Dabei können zwischen den Beteiligten Informationen über Chancen und Risiken aber auch zu Vorgehensweisen im ML-Umfeld ausgetauscht werden, um aus bisherigen Erfahrungen neue Erkenntnisse zu gewinnen. Diese Vorgehensweise wird bereits in der KI-Forschung- und Entwicklung angewendet, findet aber im Sinne der Wettbewerbsorientierung ihre Grenzen.

Eine weitere Möglichkeit, Offenheit im Kontext von ML zu fördern, besteht für alle Unternehmen gleichermaßen und unabhängig voneinander durch den Einsatz von

„Explainable-Artificial-Intelligence“ (XAI). Nachfolgend soll XAI als Lösungsansatz für Offenheit vorgestellt und kritisch beleuchtet werden.

5 Konzepte erklärbarer Künstlicher Intelligenz („Explainable Artificial Intelligence“, XAI)

Ziel von XAI ist die Schaffung von Offenheit in KI-Anwendungen durch transparente und erklärbare maschinelle Entscheidungen. Eine Erklärung stellt letztlich die finale Antwort auf eine Frage dar (Gilpin et al. 2018). Sie ist gebunden an Wissen, Kognition und an die individuellen Interpretationen eines menschlichen Empfängers. Hieraus ergibt sich ein Spannungsverhältnis zwischen Interpretierbarkeit und Vollständigkeit des Wissens. Interpretationen sind für den Rezipienten verständliche, mit für ihn bedeutungsbehafteten Begriffen formulierte Beschreibungen komplexer Systeme. Vollständigkeit fordert die Erfassung aller eine Entscheidungssituation prägenden Operationen und Informationen bzw. Parameter. Offenheit benötigt in diesem Kontext den Zugriff auf die Inputdaten des Modells und auf das trainierte ML-Modell. Mit dem Zeitpunkt, an dem die „Erklärbarkeit“ ansetzt, lassen sich drei generische XAI-Ansätze unterscheiden: (Khaleghi 2019)

- Ante-hoc Konzepte umfassen u. a. explorative Analysen von Datensätzen mittels mathematisch-statistischer Verfahren. Einblicke in zugrundeliegenden Datensätze ermöglichen hierbei z. B. Heat-Maps, Bi-Plots oder Histogramme. Ferner lassen sich statistische, datenbeschreibende Kennzahlen, wie z. B. Mittelwerte, Varianzen und Zusammenhangsmaße erfassen oder implizite Muster in den Daten z. B. durch Clusteranalysen oder andere multivariate Verfahren aufdecken. Sie helfen auch bei der Ausgestaltung einer nachvollziehbaren Dokumentation durch automatisch generierte Variablen. In diesem Kontext kann die Erfahrung von Experten helfen, Datensätze besser zu verstehen und zu erklären. Zudem können diese Verfahren im Rahmen der Entwicklung von Modellen und dem besseren Verständnis der im Unternehmen anfallenden Daten helfen.
- Design-Ansätze zielen auf die Entwicklung von einfachen, interpretierbaren Rechenmodellen ab. Dabei wird das Gesamtsystem in überschaubare Module aufgegliedert. Auf der „kleinteiligen“ modularen Ebene verspricht man sich eine bessere Erklärbarkeit („White-Box“). Solchen Modellen wird allerdings ein Zielkonflikt zwischen Erklärbarkeit und Performance („Interpretability vs. Performance Trade-off“) attestiert: Sie weisen eine Design-bedingte geringere Leistungsfähigkeit bei der Modellierung komplexer Aufgabenstellungen im Vergleich zu „Black-Box“ Modellen, wie z. B. tiefen neuronalen Netzen im Rahmen von Verfahren der Bilderkennung, auf. (Hall et al. 2020) Zusätzlich zu berücksichtigen ist, dass auch „White-Box“ Modelle sehr komplex und unübersichtlich sein können, z. B. bei hoch-dimensionalen Datensätzen, sehr tiefen Entscheidungsbäumen oder breiten Entscheidungsregeln.
- Post-hoc Konzepte lassen sich in drei Untergruppen gliedern: Techniken des Model Debugging beschäftigen sich mit Residual- und Sensitivitätsanalysen sowie mit der Generierung von adversarialen Beispielen. Im letzteren Fall kann bei-

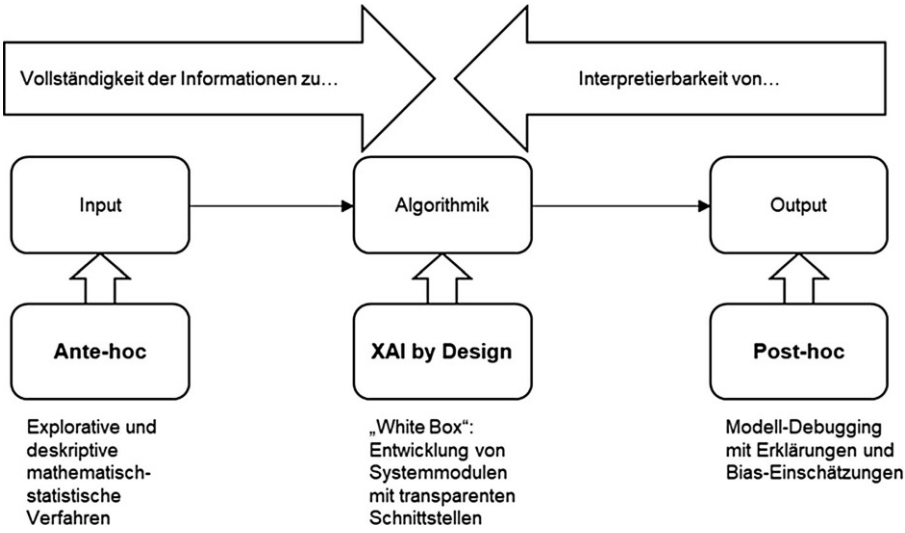

Abb. 1 Vergleich der verschiedenen XAI-Ansätze (Morelli et al. 2020, S. 11)

spielsweise die Veränderung von Pixeln eines Bildes das Modell zu einer falschen Klassifikation desselben provozieren. Demgegenüber sind Menschen häufig in der Lage, solche „Manipulationen" unmittelbar zu erkennen. Post-hoc Erklärungen kann man auf „Black-Box"-Modelle anwenden (Arrieta et al. 2020). Die Verfahren lassen sich in modell-agnostische und modell-spezifische Verfahren differenzieren. Modell-agnostische Ansätze werden im Nachgang auf ein trainiertes Modell angewendet und erzeugen eine Erklärung für die Entscheidungsfindung des Modells. Modell-spezifische Verfahren sind in ihrer Anwendungsart auf bestimmte Klassen von Modellen reglementiert. Unter Post-hoc Bias Einschätzungen lässt sich das Erkennen und Beseitigen von u. a. disparaten Einflüssen im Kontext eines Modells verstehen. Eine disparate Einflussanalyse beschreibt die Divergenz im Output für verschiedene Gruppen. So kann es bspw. basierend auf einem ML-Modell für Gruppen verschiedener Ethnien zu unterschiedlichen Kreditentscheidungen kommen. Besonders das Prozedere des sogenannten „Redlining" aus dem Finanzsektor hat hierbei Aufsehen erlangt: Auch wenn die Ethnie in den Modellen als Input-Faktor nicht explizit benutzt wird, kann es aufgrund ihrer Korrelationen mit Einkommen, Adresse und anderen Faktoren zu einer rassistisch-verzerrten Vorhersage kommen (Wang et al. 2018).

Diese generischen XAI-Konzepte lassen sich in Form von „hybriden" Ansätzen kombinieren, um den Zielkonflikt zwischen Erklärbarkeit und Performance zu lösen. Hybride Gestaltungsansätze zielen durch die Vermischung von White- und Black-Box-Modellansätzen darauf ab, die Leistungsfähigkeit bzw. die Vorhersagekraft eines Modells zu steigern, indem man z. B. ein neuronales Netz mit der Erklärbarkeit eines „White-Box" Modells kombiniert (Arrieta et al. 2020). Die Verankerung von Vorhersage und Erklärung in einem Modell setzt allerdings nachvollziehbare Trainingsdaten voraus. Die durch hybride Systeme angebotenen Erklärungen reprä-

sentieren weiterhin nicht den tatsächlichen Entscheidungspfad des Modells, sondern vielmehr eine vereinfachte, für den Menschen interpretierbare Beschreibung. Anpassungen der Architektur zur Schaffung von Erklärbarkeit fokussieren sich auf tiefe neuronale Netze und nutzen spezielle Verlust-Funktionen, um z. B. bei Prototypen einer Klassifikation zu lernen. Durch Hinzunahme von regulierenden Termen in den Verlustfunktionen sollen Modelle automatisiert ausgestaltet werden, deren Entscheidungsgrenzen sich durch „White-Box" Modelle wie Entscheidungsbäume approximieren lassen. Dieses Ersatzmodell simuliert die Vorhersage in einer verständlicheren Art.

Eine Unterteilung von XAI in Abhängigkeit von der jeweiligen Funktionsweise ist ebenfalls möglich. Hierbei haben sich vier Ansätze zur Unterteilung herauskristallisiert: pertubationsbasierte, funktionsbasierte, Surrogate-/Sampling-basierte und strukturbasierte Ansätze. Pertubationsbasierte Verfahren basieren auf dem Ansatz den Input des Models zu beeinflussen und daraufhin die Reaktion des Modells zu analysieren. Dieser Modell-agnostischer Ansatz identifiziert die Relevanz der Informationen auf Basis der Veränderung der Modellvorhersage. Meaningful Pertubations (Fong und Vedaldi 2017) ist ein Beispielansatz aus der Gruppe dieser Erklärmodelle. Funktionsbasierte Modelle, wie Sensitivity Analysis (Simonyan et al. 2014), sind den pertubationsbasierten Ansätzen sehr ähnlich, behandeln aber das zugrundeliegende ML-Modell als Funktion. Diese Methoden berechnen Quantifizierungen wie Gradienten und konstruieren Erklärungen auf Basis von diesen. Die dritte Klasse der Erklärmodelle, Surrogate-/Sampling-basierte Ansätze, sind ebenfalls modell-agnostische Ansätze, wie bspw. LIME (Ribeiro et al. 2016). Sie beruhen auf dem Ansatz lokaler Annäherungen durch simple Funktionsmodelle und bilden Erklärungen auf dieser Basis. Strukturbasierte Ansätze, wie Layerwise Relevance Propagation (Bach et al. 2015), zeichnen sich durch eine modell-spezifische Funktionsweise aus. Diese berücksichtigen die Struktur des zugrundeliegenden ML-Modells und teilen bei komplexen Funktionen die Erklärung in Subfunktionen auf, um diese zu erklären und zu aggregieren.

Die aufgezeigten XAI-Ansätze bieten einerseits viele Möglichkeiten, Wissen über trainierte Modelle zu erlangen. Andererseits sind aber auch deren Grenzen näher zu definieren. Der Zielkonflikt zwischen Interpretierbarkeit und Vollständigkeit sorgt in komplexen Systemen leicht für Situationen, in denen das System entweder zu komplexe oder zu einfache und dadurch letztendlich irreführende Erklärungen produziert. Abstrakte Vektorräumen sind für den Menschen nicht verständlich und damit auch nicht interpretier- und erklärbar. (Holzinger et al. 2018) Post-hoc Ansätze können die Entscheidungsgrenzen der Modelle schätzen und erklären, aber sie berechnen keine optimale Lösung. Zudem zeigen die dargestellten XAI-Ansätze zwar Überschneidungen mit den aktuell diskutierten Ansätzen zur „Fairness-Schätzung" (Gajane und Pechenizkiy 2018), sind damit aber nicht gleichzusetzen. Bei diesen Ansätzen geht es im Kern darum, Verzerrungen in den für das Training verwendeten Datensätzen und damit verbundene Diskriminierungen auf unterschiedlichen Ebenen in maschinellen Entscheidungen erkennen und ausschließen zu können. XAI-Methoden liefern eine Hilfestellung dafür, Verzerrungen in den Trainingsdaten transparent sowie erklärbar zu machen und leisten insoweit einen Beitrag zur Schaffung von „Fairness". Sie können Modell-Artefakte und versteckte Korrelationen erkennbar machen und

es somit den Entwicklern ermöglichen den Grund für mögliche Diskriminierungen erkennbar zu machen. (Arrieta et al. 2020) Es ist daher erforderlich, Varianten der Fairnessschätzung mit XAI-Methoden zu kombinieren.

Eine Bewertung der berechneten Erklärungen sollte im unternehmerischen Kontext der Anwendung erfolgen und kann anhand verschiedener Kriterien bspw. durch das „Fact-Sheet“ vorgenommen werden. (Sokol und Flach 2020) Neben funktionalen und operativen Anforderungen stellen Nutzbarkeit, Sicherheit und Privatsphäre sowie Validierungen der Erklärungen durch die Nutzer verschiedene Bewertungsgruppen mit jeweils mehreren Kriterien dar. Die Kausalität einer Erklärung, als Teil der operativen Anforderungen, stellt dabei eine besondere Herausforderung dar. Dem Nutzer muss bewusst sein, dass ein Unterschied zwischen Erklärungen aus kausalen- und nicht-kausalen Modellen besteht, da ansonsten kausale Fehlschlüsse auftreten. Eine Vertiefung dieser Bewertung kann im Kontext der Kausalität anhand des 3-Schichten kausalen Hierarchie-Modells (Pearl 2018) analog der Forderung nach „Causability“ (Holzinger et al. 2019) vorgenommen werden. Der Ausdruck „Causability“ stellt hierbei den Grad des geschaffenen kausalen Verständnisses einer Erklärung dar, bewertet anhand von Effektivität, Effizienz und Nutzerzufriedenheit. Das 3-Schichten-Modell zur kausalen Hierarchie beschäftigt sich auf der ersten Schicht mit der Beeinflussung des Glaubens über den Modelloutput durch die Beobachtungen eines Einflussfaktors. Dabei soll die Frage nach dem „Wie verändert sich mein Glaube über das Eintreten von Y wenn ich X beobachte?“ geklärt werden. Auf zweiter Schicht erfolgt eine Bewertung anhand des Einflusses im Rahmen der Durchführung eines bestimmten den Output beeinflussenden Faktors, also die Beantwortung der Frage: „Was passiert, wenn eine bestimmte Tätigkeit X durchgeführt wird?“. Im Rahmen der dritten Schicht wird die Retrospektive verwendet um eine Einschätzung über die Ursachen-Wirkungsbeziehung bei bereits beobachteten Daten herausstellen zu können.

Diese Schichten sollten anhand der Messung zur Effektivität der Erklärung, der Effizienz im Sinne von minimaler Zeit und Aufwand, sowie der Nutzerzufriedenheit bewertet werden. Die Evaluation der Nutzerzufriedenheit ist bereits auf der Ebene des „Fact-Sheets“ integriert, wird aber nicht im Rahmen des kausalen Hierarchie-Modells herangezogen und könnte somit um die Bewertung des kausalen Zusammenhangs erweitert werden. Als Maß zur Berechnung der Nutzerzufriedenheit kann beispielsweise die „System Causability Scale“ (Holzinger et al. 2020), basierend auf der Likert Skala, herangezogen werden.

Der Bedarf an Offenheit im Kontext von KI-basierten Entscheidungen entsteht daraus, dass ohne eine Erklärbarkeit der Entscheidungsfindung, keine Möglichkeit besteht diesen zu verifizieren, zu verbessern oder von diesem zu lernen (O’Neil 2016). XAI als Tool zur Verdeutlichung der Entscheidungsfindung von ML-gestützten Entscheidungen ist dabei bereits integraler Bestandteil im KI-Entwicklungsvorgehen von Unternehmen.

6 XAI – ein Schlüssel zur Verbesserung von ML-unterstützten Unternehmensentscheidungen?

Um rationale unternehmerische Entscheidungen zu fördern und diese validierbar zu machen, ist der Mittel-Zweck-Charakter von Entscheidungen im Unternehmen zu reflektieren. Sieht man ein Unternehmen als sozioökonomisches System an, muss die Frage gestellt werden, welche Ziele verfolgt werden und in welchem Verhältnis diese zueinanderstehen. Dabei können auch „verdeckte", d. h. nicht offen gelegte Zielsetzungen existieren. Rationalität basiert darüber hinaus auf Vollständigkeit der Entscheidungsalternativen, die trotz der Verfügbarkeit von großen Datenmengen nicht gegeben sein muss.

ML-Systeme bieten Chancenpotenziale, unternehmerische Entscheidungen durch geeignete Datenmodelle zu verbessern. Die damit verbundenen Risiken sowie regulatorische Aspekte machen Offenheit im Sinne eines XAI-Einsatzes erforderlich, um entsprechende Vorhaben zu fördern. Entscheidungen müssen intersubjektiv nachvollziehbar begründet werden und ein Mindestmaß an Valenzen im soziologischen Sinn und an Resilienz aufweisen. Ferner sind Erkenntnisse aus Untersuchungen von Gruppenentscheidungen zu berücksichtigen, um die Interaktion zwischen Benutzer(n) und ML-Systemen adäquat auszugestalten. Die aufgezeigte Problematik der XAI-Verfahren verdeutlicht die Notwendigkeit zusätzlicher Instrumente zur Schaffung von Transparenz und Rationalität in unternehmerischen Entscheidungen.

Unter Heuristiken versteht man generell aufgestellte Regeln zur Entwicklung von „guten" Lösungen, ohne deren Optimalität und Zuverlässigkeit garantieren zu können („Daumenregeln"). Bei der Konstruktion und Evaluation kommt es primär auf deren praktische Einsetzbarkeit und Lösungsqualität an. Entscheidendes Prüfkriterium für die Beurteilung von Heuristiken im XAI-Kontext ist die Frage: Sind sie zur Konstruktion von kooperativen Entscheidungssituationen zwischen Mensch und Maschine geeignet? Dies ist der Fall, wenn die zusätzlich generierte Offenheit die Ableitung besserer Lösungen in unternehmerischen Entscheidungssituationen ermöglicht.

In Analogie zum Ansatz der Evaluierung von Industrie 4.0 Konzepten als soziotechnische Systeme („Heuristik 4.0") (Hirsch-Kreinsen und Karačić 2019) wird ein sich auf mehrere Ebenen erstreckendes Postulat für die Ausgestaltung der Kooperation vorgestellt:

- *Unterstützungstechnologie*: Als Vorgabe für die Interaktion hat das ML-System dem menschlichen Entscheidungsträger rechtzeitig und in geeigneter Form Vorschläge zu unterbreiten und keine Anweisungen zu tätigen. Der Entscheidungsprozess darf durch das ML-System nicht behindert werden. Das System steht im Rahmen einer Entscheidungsunterstützung für eine Arbeitsteilung, in welcher der Mensch bei auftretenden Unsicherheiten maßgeblich die Verantwortung trägt.
- *Nachvollziehbarkeit*: Bei der Abwägung zwischen der Modellperformanz (z. B. Genauigkeit, Präzision) und Modell-Deutbarkeit ist die zweite Kategorie für ein kooperatives Zusammenwirken bedeutsamer. Entsprechend muss ein XAI-Interaktionssystem Bestandteil des ML-Systems sein. Die Transparenz von ML-Syste-

men muss auf unterschiedlichen Ebenen verfügbar sein (Gesamtmodell, einzelne Komponenten und Algorithmen-Einsatz).

- *Flexibilität*: Die Berücksichtigung der Ergebnisse durch ML-Systeme obliegt den menschlichen Entscheidungsträgern. Beachten diese nicht die maschinellen Vorschläge, spricht dies für eine erneute Evaluierung und ggf. Weiterentwicklung des ML-Systems. Aus XAI-Perspektive kommt in diesem Zusammenhang insbesondere Ante-hoc-Konzepten eine besondere Bedeutung zu, da diese Beschreibungen der vorhandenen Daten ermöglichen und somit helfen menschliches Expertenwissen adäquat zu berücksichtigen. Analog des CRISP-DM Entwicklungsvorgehens, können Verfahren der Zusammenhangsanalyse oder Cluster-Analysen helfen ein Verständnis zwischen dem Daten- und Geschäftsverständnis zu bilden.
- *Kommunikationsunterstützung*: Das methodisch (Vor-)Wissen und mentale Modelle der Entscheidungsträger und der Stakeholder sind bei der Ausgestaltung zu berücksichtigen. Generell sind visuelle Erklärungen oder Erklärungen in natürlicher Sprache bei der Ausgestaltung der Mensch-ML-Schnittstelle zu präferieren. Die Entscheidungsfreiheit, ob und wann Erklärungen abgerufen werden, liegt primär bei den menschlichen Nutzern und sollte regelmäßig evaluiert und durch die Nutzer anpassbar sein. (Sokol und Flach 2020)
- *Informationsaustausch*: Bereits im Vorfeld bei der Datengenerierung ist zu gewährleisten, dass diese den Unternehmens-Governance- und -Compliance-Anforderungen genügen. Die Nutzer werden über die Daten, welche das System verarbeitet, sowie über deren Priorisierung in Kenntnis gesetzt. Zudem sind der durch das System berechnete Zielwert sowie abgeleitete Konsequenzen an die Nutzer zu kommunizieren. Ferner sind die bereitgestellten Daten in geeigneten Zeitabständen regelmäßig zu aktualisieren.
- *Balance*: Im Rahmen der Kooperation ist einerseits die Zielsetzung der Entscheidungsträger zu berücksichtigen. Andererseits ist die Wirksamkeit des ML-Modells mit entsprechend großer Datengrundlage Rechnung zu tragen. Weiterhin ist der Aufbau von Erfahrungen bei den beteiligten Mitarbeitern durch das Unternehmen in geeigneter Weise zu fördern. Hierzu lassen sich Post-hoc Konzepte in Form eines Modell-Debugging mit Erklärungen gezielt einsetzen.
- *Kompatibilität*: Die erzielten Kooperationsergebnisse sollen den Wertorientierungen der Unternehmensleitung entsprechen. Die zugehörige Zielstruktur ist im Entscheidungsmodell abzubilden und muss konsistente Ergebnisse liefern. Die Auswahl der Methoden im Rahmen eines XAI-Systems muss dabei anhand vordefinierter Kriterien erfolgen (Sokol und Flach 2020) und durch geeignete Prüfungen regelmäßig erneut evaluiert werden. In diesem Kontext ist darauf zu achten, dass die von den ML-Algorithmen gelernten Artefakte keine Diskriminierungen enthalten bzw. Fairnessaspekte hinreichend berücksichtigt werden.
- *Organisationseffizienz*: Der Entscheidungsprozess ist möglichst durchgängig durch XAI-Systeme zu unterstützen. Dies spricht für den Einsatz von hybriden XAI-Systemen im Rahmen der Modellauswahl. Weiterhin ist ein systematischer Kompetenzaufbau (Sokol und Flach 2020) auf Seiten der Nutzer für den Umgang mit XAI-Systemen zu betreiben, wobei eine regelmäßige Validierung des XAI-Systems durch Entscheidungsträger z. B. im Kontext des in Kap. 5 beschriebenen

„Fact-Sheets“ erfolgen kann. Ferner ist der Situation Rechnung zu tragen, dass es sich um dynamische, lernende Systeme handelt.

Das Ziel besteht in der Identifikation möglicher Verbesserungspotenziale für konkrete ML-Projekte zur Unterstützung operativer betriebswirtschaftlicher Entscheidungen. Der Wirkungsgrad einer Heuristik lässt sich allerdings erst bestimmen, wenn hinreichende Einsatzerfahrungen vorliegen. Sowohl die aufgezeigten Kriterien als auch der Austausch von Erkenntnissen im praktischen Einsatz basieren auf offenem Feedforward und Feedback. Dieses Konstrukt erweist sich als Grundlage für einen verantwortungsvollen Umgang mit dem Einsatz von ML für unternehmerische Entscheidungen. Offenheit kann damit einen wichtigen Beitrag für innovative Weiterentwicklungen im XAI-Bereich führen.

Funding Open Access funding enabled and organized by Projekt DEAL.

Literatur

Alizadeh F, Margarita E, Stevens G (2020) eXplainable AI: take one step back, move two steps forward. Workshop on User-Centered Artificial Intelligence (UCAI ’20).

Alliance EA (2020) ALTAI – the assessment list on trustworthy artificial intelligence. https://futurium.ec.europa.eu/en/european-ai-alliance/pages/altai-assessment-list-trustworthy-artificial-intelligence. Zugegriffen: 19.02.2021

Amann K, Petzold J (2014) Management und Controlling – Instrumente – Organisation – Ziele. Gabler, Wiesbaden

Arrieta AA, Díaz-Rodríguez N, Del Ser J, Bennetot A, Tabik S, Barbado A (2020) Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. Inf Fusion 58(2020):82–115

Bach et al. (2015). On Pixel-Wise Explanations for Non-Linear Classifier Decisions by Layer-Wise Relevance Propagation. Von https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0130140 abgerufen

Bughin J, Seong J, Manyika J (2018) Notes from the AI frontier: Modeling the impact of AI. World Econ. McKinsey Global Institute, New York

Chapman P, Clinton J, Kerber R, Khabaza T, Reinartz T, Shearer C, Wirth R (2000) CRISP. Dis Mon. 1.0 Step-by-step data mining guide. The CRISP-DM consortium

Chlupsa C (2017) Der Einfluss unbewusster Motive auf den Entscheidungsprozess. Wie implizite Codes Managemententscheidungen steuern. https://doi.org/10.1007/978-3-658-07230-8. Springer Gabler, Wiesbaden

Clark A (2018) The Machine. Learning. Audit—CRISP-DM Framework. Rolling Meadows: ISACA

Dastin J (2018). Amazon scraps secret AI recruiting tool that showed bias against women. Von Reuters: https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G. Zugegriffen: 19.02.2021
European Commission. (25. April 2018). Communication Artificial Intelligence for Europe. Von https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=51625. Zugegriffen: 19.02.2021
Fong R, Vedaldi A (2017) Interpretable explanations of black boxes by meaningful perturbation. Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV). Venice, Italy: IEEE. https://arxiv.org/abs/1704.03296. Zugegriffen: 19.02.2021
Gajane, P., & Pechenizkiy, M. (28. 05 2018). On Formalizing Fairness in Prediction with Machine Learning. Von https://arxiv.org/abs/1710.03184. Zugegriffen: 19.02.2021
Gilpin LH, Bau D, Yuan BZ, Bajwa A, Specter M, Kagal L (2018) Explaining Explanations: An Overview of Interpretability of Machine Learning. 5th IEEE International Conference on Data Science and Advanced Analytics (DSAA 2018). Turin, Italy: IEEE. https://arxiv.org/abs/1806.00069. Zugegriffen: 19.02.2021
Goodfellow I, Bengio Y, Courville A (2016) Deep Learning. MIT Press, Cambridge
Haenssle HA, Fink C, Schneiderbauer R, Toberer F, Buhl T, Blum A et al (2018) Man against machine: diagnostic performance of a deep learning convolutional neural network for dermoscopic melanoma recognition in comparison to 58 dermatologists. Ann Oncol 29(8):1836–1842
Hall P, Gill N, Kurka M, Phan W (2020) Machine learning interpretability with H2O driverless AI. http://docs.h2o.ai/. Zugegriffen: 19.02.2021
Hirsch-Kreinsen H, Karačić A (2019) Digitalisierung von. Arbeit, Bd. 16. Forschungsinstitut fur gesellschaftliche Weiterentwicklung (e. V.), Düsseldorf
HLEG-AI. (07. 01 2021). A definition of AI: Main capabilities and scientific disciplines. Von https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=56341. Zugegriffen: 19.02.2021
Holtmann JP (2008) Pfadabhängigkeit strategischer Entscheidungen. Gabler, Wiesbaden
Holzinger A, Langs G, Denk H, Zatloukal K, Müller H (2018) Causability and explainability of artificial intelligence in medicine. https://doi.org/10.1002/widm.1312
Holzinger A, Langs G, Denk H, Zatloukal K, Müller H (2019) Causability and explainability of artificial intelligence in medicine. WIREs Data Mining and Knowledge Discovery 9(4)
Holzinger A, Carrington A, Müller H (2020) Measuring the quality of explanations: the system causability scale (SCS). KI – Künstliche Intelligenz 34(2):193–198
COBIT (2019) by ISACA. Framework: Governance and Management Objectives.
Khaleghi B (2019) The how of explainable AI: pre-modelling explainability. https://towardsdatascience.com/the-how-of-explainable-ai-pre-modelling-explainability-699150495fe4. Zugegriffen: 19.02.2021
Laux H, Gillenkirch RM, Schenk-Mathes HY (2018) Entscheidungstheorie. 10. Aufl. ISBN 978-3-662-57817-9, ISBN 978-3-662-57818-6 (eBook). Springer, Berlin
Lipton ZC (2016) The Mythos of Model Interpretability. ICML Workshop on Human Interpretability in Machine Learning (WHI 2016), New York
Lossos C, Morelli F, Geschwill S (2019) Entwicklung einer Methodik zur Prüfung der Wirksamkeit von künstlicher Intelligenz. Arbeitskreis Wirtschaftsinformatik an Hochschulen für angewandte Wissenschaften
Mehrabi N, Morstatter F, Saxena N, Lerman K, Galstyan A (2019) A Survey on Bias and Fairness in Machine Learning. https://arxiv.org/abs/1908.09635. Zugegriffen: 19.02.2021
Merriam-Webster (2021) Definition of machine learning. https://www.merriam-webster.com/dictionary/machine%20learning. Zugegriffen: 19.02.2021
Merriam-Webster (2020) Definition of artificial intelligence. Von https://www.merriam-webster.com/dictionary/artificial%20intelligence. Zugegriffen: 19.02.2021
Morelli F, Geschwill S, Zerr K, Lossos C (2020) Rationalität maschineller Entscheidungen im Unternehmen. Tagungsband zur 33. Jahrestagung des Arbeitskreises Wirtschaftsinformatik der deutschsprachigen Fachhochschulen (AKWI)
Neifer T, Lawo D, Esau M (2021) Data science canvas: evaluation of a tool to manage data science projects. Hawaii International Conference on System Sciences 2021
O'Neil C (2016) Weapons of Math Destruction: How Big Data Increases Inequality and Threatens. Crown, New York City
Pearl J, Mackenzie D (2018) The Book of Why: The New Science of Cause and Effect. 1. Aufl. Basic Books, New York
Ribeiro MT, Singh S, Guestrin C (2016) "Why Should I Trust You?": Explaining the Predictions of Any Classifier. https://arxiv.org/abs/1602.04938. Zugegriffen: 19.02.2021
Russell SJ, Norvig P (2016) Artificial Intelligence: A Modern Approach. Harrow. Pearson, UK

Sartor G, Lagioia F (2020) The impact of the Genereal Data Protection Regulation (GDPR) on artificial intelligence. https://www.europarl.europa.eu/RegData/etudes/STUD/2020/641530/EPRS_STU(2020)641530_EN.pdf. Zugegriffen: 19.02.2021

Schencking F (2018) Rationalität des Entrepreneurs versus Rationalität des Managers. In G. Faltin, Handbuch Entrepreneurship. Gabler, Wiesbaden

Schlagwein D, Conboy K, Feller J, Leimeister JM, Morgan L (2017) "Openness" with and without information technology: a framework and a brief history. J Inform Tech 32(4):297–305

Simonyan K, Vedaldi A, Zisserman A (2014) Deep inside convolutional networks: visualising image classification models and saliency maps. https://arxiv.org/abs/1312.6034. Zugegriffen: 19.02.2021

Sokol K, Flach P (2020) Explainability Fact Sheets: A Framework for Systematic Assessment of Explainable Approaches. Conference on Fairness, Accountability, and Transparency. FAT*, Bd. 20. ACM, Barcelona, Spain

Verma S, Rubin J (2018) Fairness definitions explained. 2018 IEEE/ACM International Workshop on Software Fairness (FairWare)

Wang H, Ustun B, Calmon FP (2018) On the direction of discrimination: an information-theoretic analysis of disparate impact in machine learning. https://arxiv.org/pdf/1801.05398. Zugegriffen: 19.02.2021