Fachhochschule Graubünden
University of Applied Sciences

Churer Schriften
zur Informationswissenschaft

Herausgegeben von
Wolfgang Semar, Bernard Bekavac, Ivo Macek

Arbeitsbereich Bachelor of Science
in Digital Business Management

Schrift 180

Genderungleiche KI-Anwendungen im Recruiting

Analyse und Erarbeitung von Massnahmen, um einen Geschlechterbias durch den Einsatz von Künstlicher Intelligenz im Recruiting zu vermeiden.

Manuel Fercher

Chur 2024

Churer Schriften zur Informationswissenschaft

Herausgegeben von Wolfgang Semar,
Bernard Bekavac, Ivo Macek

Schrift 180

Genderungleiche KI-Anwendungen im Recruiting

Analyse und Erarbeitung von Massnahmen, um einen Geschlechterbias durch den Einsatz von Künstlicher Intelligenz im Recruiting zu vermeiden.

Manuel Fercher

Diese Publikation entstand im Rahmen einer Thesis zum Bachelor of Science in Digital Business Management.

Referent: Dr. Vincenzo Francolino

Korreferent: Urban Kalbermatter

Verlag: Fachhochschule Graubünden

ISSN: 1660-945X

Ort, Datum: Chur, Dezember 2024

Abstract

Die vorliegende Bachelorthesis beschäftigt sich mit der Frage, welche Massnahmen ergriffen werden können, um der Diskriminierung von Frauen durch Künstliche Intelligenz in der Personalrekrutierung entgegenzuwirken. Dazu wurde eine Literaturrecherche durchgeführt, in der der aktuelle Einsatz von KI-Tools im Recruiting sowie die Technologie an sich näher betrachtet wurden. Dabei zeigte sich, dass KI bereits für verschiedene Aufgaben des Recruitings eingesetzt wird und dabei Bias auftreten können, die Frauen benachteiligen. Sowohl für die Entwicklung als auch für den Einsatz solcher KI-Tools wurden Handlungsempfehlungen definiert. Um einen diskriminierungsfreien Einsatz von Tools zu gewährleisten, werden Massnahmen aufgezeigt, die vor, während und nach der Entwicklung solcher Systeme ergriffen werden können. Da viele dieser Massnahmen eine erklärbare KI voraussetzen, wird dieser Ansatz als wesentliches Element dieser Arbeit behandelt.

Inhaltsverzeichnis

1 Einleitung 1

1.1 Forschungsfrage 2
1.2 Methodisches Vorgehen 3

2 Künstliche Intelligenz im Recruiting 7

2.1 Anforderungen an das Recruiting 7
2.2 Entwicklung zum Data Driven Recruiting 8
2.3 Aktueller Einsatz von Künstlicher Intelligenz 9
2.4 Gründe für den Einsatz von Künstlicher Intelligenz im Recruiting 11
2.5 Empfehlungen für die HR-Abteilung aus Whitepapers 12
2.6 Beispiele von aufgetretenem Genderbias 13
2.7 Kritik an KI-Tools im Recruiting 14
2.7.1 Einstellung der Bewerbenden 15
2.7.2 Einstellung der Recruiter 15

3 Funktionsweise Künstlicher Intelligenz 17

3.1 Einteilung der KI 17
3.1.1 Neural Networks 18
3.1.2 Machine-Learning 19
3.1.3 Deep Learning 22
3.2 Blackbox-Problem 22
3.3 Bias 24
3.3.1 Dateninduzierter Bias 24
3.3.2 Modellinduzierter Bias 25
3.3.3 Weitere Bias 27

4 Massnahmen gegen einen Geschlechterbias 29

4.1 Erklärbare KI (Whitebox-Ansatz) 29
4.1.1 Ablauf der XAI-Entwicklung 30
4.1.2 Blackbox vs. Whitebox 32
4.2 Pre-processing Massnahmen 34
4.3 In-processing Massnahmen 35
4.4 Post-processing Massnahmen 35
4.5 Laufende Massnahmen 36
4.5.1 Zusammenarbeit zwischen Menschen und KI definieren 36

4.5.2 Festlegen spezifischer Anwendungsfälle 37
4.5.3 Ethische Prinzipien und Richtlinien 38
4.5.4 Schulungen und Awareness 39
4.5.5 Transparenz gegenüber Bewerbenden schaffen 40
4.5.6 Kontinuierliche Überwachung 40

5 Diskussion 41

5.1 Interpretation der Ergebnisse 41
5.1.1 Beantwortung Unterfrage 1 41
5.1.2 Beantwortung Unterfrage 2 42
5.1.3 Beantwortung Forschungsfrage 42
5.2 Limitationen und Stärken 45

6 Fazit 47

7 Literaturverzeichnis 49

Abbildungsverzeichnis

Abbildung 1: Phasen des Recruitings, (Eigene Darstellung in Anlehnung an Z. Chen, 2023) 9
Abbildung 2: Einordnung Künstliche Intelligenz, Machine-Learning und Deep Learning, (Eigene Darstellung angelehnt an Wuttke, 2023) 17
Abbildung 3: Neural Networks, (Eigene Darstellung angeleht an Dike et al., 2018, S.322; Suzuki, 2011, S. 6) 18
Abbildung 4: Supervised Learning, (Eigene Darstellung angelehnt an Dike et al., 2018, S.323) 20
Abbildung 5: Unsupervised Learning, (Eigene Darstellung angelehnt an Dike et al., 2018, S. 324) 21
Abbildung 6: Reinforcement Learning, (Eigene Darstellung angelehnt an Dike et al, 2018, S.323) 22
Abbildung 7: Input-Output-Beziehung im Blackbox-Modell, (Eigene Darstellung) 23
Abbildung 8: Ursachen für einen Bias in Algorithmen, (Eigene Darstellung angelehnt an Langer und Weyerer, 2020, S. 224) 27
Abbildung 9: Ablauf Erklärbare KI, (Eigene Darstellung angelehnt an Dwivedi, 2023, S. 3-5) 31
Abbildung 10: Blackbox vs. Whitebox, (Eigene Darstellung) 32
Abbildung 11: Trade-off zwischen Genauigkeit und Erklärbarkeit bei KI-Modellen, (Eigene Darstellung angelehnt an Bar-redo Arrieta et al., 2020, S. 31; Gunning und Aha, 2019, S. 46) 33
Abbildung 12: Entscheidungsframework durch XAI, (Eigene Darstellung angelehnt an Gunning und Aha, 2019, S. 50) 37
Abbildung 13: Vergleich ethischer Prinzipien, (Eigene Darstellung angelehnt an Barton und Pöppelbuss, 2022. S. 476-479; Fjeld et al., 2020, S.5) 39
Abbildung 14: Zusammengefasste Massnahmen für die Entwicklung und die Personalrekrutierung, (Eigene Darstellung) 45

1 Einleitung

Die zunehmende Integration von Künstlicher Intelligenz (KI) in verschiedenen Lebensbereichen hat die Art und Weise des Lebens und Arbeitens grundlegend verändert. Insbesondere die Einführung von Tools wie ChatGPT von Open AI hat die Technologie für die breite Masse zugänglich gemacht. In den letzten Jahren haben sich dadurch zahlreiche Jobs massgeblich verändert. So werden auch Aufgaben im Recruiting zunehmend von KI übernommen bzw. werden Recruiter durch KI in ihrer Arbeit unterstützt (Adelmann & Wiedmer, 2017, S. 1). Hierbei kann KI Personalabteilungen unterstützen und in einem umkämpften Wettbewerb um Arbeitskräfte einen Wettbewerbsvorteil schaffen (Guenole & Feinzig, 2018). Dabei können KI-Tools das Fachpersonal der Human Resources (HR) über den gesamten Bewerbungsprozess hinweg unterstützen und bei der Entscheidungsfindung helfen (Böhm et al., 2021, S. 195–198; Pohlink & Fischer, 2021, S. 156).

Seit der Digitalisierung Ende der 1990er Jahre hat sich der Bewerbungsprozess stark vereinfacht, was zu einer Zunahme in der Anzahl der Bewerbungen geführt hat. Dadurch sind Unternehmen zunehmend dazu gezwungen, KI zur Bewältigung der hohen Datenmengen im Recruiting einzusetzen (Black & van Esch, 2020, S. 217; Jares & Vogt, 2021, S. 75). Einer der grössten Vorteile von KI im Recruiting ist die Effizienzsteigerung, die zu einer erheblichen Kostenreduktion führt (Black & van Esch, 2020, S. 222; Wilke & Bendel, 2022, S. 655).

Aufgrund des Blackbox-Charakters mancher KI besteht die Gefahr, dass der Entscheidungsprozess von Algorithmen für menschliche Entscheidungsträgerinnen und Entscheidungsträger nicht mehr nachvollziehbar ist, da die Kausalität verlorengeht, auch wenn das Ergebnis der KI korrekt ist (Heim & Gerth, 2023, S. 266). So ist es möglich, dass nicht beurteilt werden kann, warum eine KI bestimmte Entscheidungen im Bewerbungsprozess so getroffen hat (Böhm et al., 2021, S. 198). Ein anderes grosses Problem ist, dass KI mögliche Verhaltensmuster und Denkweisen der Gegenwart reproduziert. Somit kann es vorkommen, dass durch solche Tools eine Diskriminierung gegenüber bestimmten Personengruppen stattfindet (Böhm et al., 2021, S. 199). Ein bekanntes Beispiel, bei dem ein geschlechtsspezifischer Bias (Verzerrung) Frauen systematisch benachteiligte, bietet Amazon. Das Unternehmen musste ein KI-Rekrutierungstool einstellen, da es Frauen für bestimmte Berufe, insbesondere in technischen Bereichen, systematisch schlechter einstufte als Männer. Diese Voreingenommenheit war darauf zurückzuführen, dass die Musterdaten für eine «passende» bewerbende Person hauptsächlich von Männern stammten. Der Algorithmus schloss daraus, dass Frauen für diese Berufe weniger geeignet seien (Lavanchy, 2018).

In dieser Bachelorthesis wird untersucht, wie ein solcher geschlechtsspezifischer Bias verhindert werden kann. Dazu wird in einem ersten Schritt in Kapitel 3 analysiert, wie KI derzeit in der Personalrekrutierung eingesetzt wird. Des Weiteren wird in diesem Kapitel die Entwicklung des Recruitings in den letzten dreissig Jahren aufgezeigt, da diese relevant ist, um zu verstehen, warum KI aktuell eingesetzt wird. Ausserdem werden verschiedene Fälle von aufgetretener Geschlechterdiskriminierung durch KI im Recruiting erläutert. In Kapitel 4 werden die Technologie der KI und deren Unterbegriffe Machine-Learning und Deep Learning erklärt. Dabei wird aufgezeigt, wo und warum bei der Technologie ein Bias entstehen kann. Anschliessend werden im Kapitel 5 Massnahmen aufgezeigt, um zukünftig einen Bias in der Personalrekrutierung vermeiden zu können. Diese Massnahmen bilden die Grundlage für die Beantwortung der Forschungsfrage in Kapitel 6.1.

Die Relevanz des Themas ergibt sich aus dem Einsatz von KI und den damit verbundenen Risiken. Zum einen birgt der Einsatz von KI im Recruiting die Gefahr einer Diskriminierung gegenüber Personengruppen, zum anderen besteht die Gefahr, dass KI-Systeme durch eine Blackbox nicht nachvollziehbar sind. Dies erschwert es, die Fairness der Entscheidungen zu überprüfen und gegebenenfalls einzugreifen. Diese Problematik wird auch in verschiedenen Whitepapers aus der Recruiting-Branche erwähnt (Employ & JOBVITE, 2023; Onlyfy, 2023; Personio, 2023a). Angesichts dieser Herausforderung ist es von grosser Bedeutung, Massnahmen zu entwickeln und zu implementieren, die einen fairen und diskriminierungsfreien Einsatz von KI im Recruiting gewährleisten. Dies kann sowohl technische Massnahmen für die Entwicklung solcher Systeme als auch Massnahmen für den Umgang mit solchen Systemen umfassen.

1.1 Forschungsfrage

Wie bereits erläutert, ist die Problematik des Einsatzes von KI-Tools im Recruiting mittlerweile weitgehend bekannt. Für den Umgang mit solchen Anwendungen werden daher Handlungsempfehlungen benötigt. Dies betrifft einerseits die Entwicklerinnen und Entwickler solcher Tools und andererseits Anwenderinnen und Anwender in der Personalrekrutierung. Daraus leitet sich die folgende Forschungsfrage ab:

Welche Massnahmen können ergriffen werden, um einen Geschlechterbias durch den Einsatz von KI-Tools im Recruiting zu vermeiden?

Um die Forschungsfrage genauer beantworten zu können, wurden zwei Unterfragen definiert, welche ins Thema einleiten sollen. Die Unterfragen lauten folgendermassen:

Wie wird KI aktuell bei Unternehmen im Recruiting-Prozess eingesetzt?

Diese Frage zielt darauf ab, einen Überblick über die aktuellen Anwendungen und Praktiken von KI in der Personalbeschaffung zu geben. Es wird untersucht, in welchen Phasen des Bewerbungsprozesses KI-Tools eingesetzt werden und welche Vorteile und Herausforderungen damit verbunden sind.

Wie funktionieren KI-Tools, sodass dabei ein Geschlechterbias entsteht?

Diese Frage konzentriert sich auf die technischen Grundlagen der KI-Technologie. Es wird analysiert, wie und warum ein Bias in Algorithmen entstehen kann und welche spezifischen Prozesse dazu führen, dass bestimmte Gruppen, insbesondere Frauen, benachteiligt werden.

1.2 Methodisches Vorgehen

Um die Forschungsfragen beantworten zu können, wird eine Literaturrecherche durchgeführt. Diese erfolgt über die Plattformen Google Scholar, Springer Link und ScienceDirect. Die Literaturrecherche findet einerseits systematisch und andererseits unsystematisch statt. Zunächst wird systematisch nach Suchbegriffen gesucht, die nach Döring und Bortz (2016, S. 158) in primäre und sekundäre Suchbegriffe eingeteilt werden. Dabei werden folgende Schlagwörter abgefragt:

- **Primäre Suchbegriffe**: Künstliche Intelligenz, Funktionalität, Ethik, Algorithmus, Algorithmen, automatisiert, Gefahren, Risiken, maschinelles Lernen, tiefes Lernen und Blackbox
- **Sekundäre Suchbegriffe (Phase 1)**: Personalrekrutierung, HR, Tools, Rekrutierungstools, Anforderungen, Anwendungsfälle, Software, Guidelines, Handbuch, Personalsuche und Bewerbung
- **Sekundäre Suchbegriffe (Phase 2)**: Bias, Geschlechterbias, Verzerrung, frauenfeindlich, Entstehung, Gründe, sexistisch, Diskriminierung, Fälle und Beispiele
- **Sekundäre Suchbegriffe (Phase 3)**: Empfehlungen, Möglichkeiten, Fair, Gerecht, Anti-Bias, erklärbare KI, Richtlinien, Lösungen, XAI, Konzepte, erklärbare KI, Whitebox, Interpretierbarkeit, Massnahmen, Modelle und Ablauf

Die Liste der Suchbegriffe ist nicht abschliessend und wird ergänzt durch Synonyme, Kombinationen oder Abkürzungen (z. B. «KI») der Begriffe. Die Abfragen werden zudem auf Englisch durchgeführt, da dadurch eine umfassendere aktuelle Literatur zur Verfügung steht. Ausserdem wird die Suche mit Operatoren wie AND oder OR eingegrenzt, damit die Treffer relevanter und spezifischer sind (Döring & Bortz, 2016, S. 160). Zahlreiche Suchbegriffe sind zu weit gefasst und generieren allein keine relevanten Treffer.

Beispielsweise könnten die Suchanfragen so aussehen:

- Geschlechterbias AND Künstliche Intelligenz AND Personalrekrutierung
- (Gründe OR Entstehung OR Erklärungen) AND Bias AND Künstliche Intelligenz
- (Empfehlungen OR Möglichkeiten OR Lösungen) AND (faire OR gerechte OR diskriminierungsfreie) AND Künstliche Intelligenz AND (Personalsuche OR Rekrutierung OR Bewerbung)

Parallel dazu wird aufgrund dieser ersten Quellen eine unsystematische Literatursuche durchgeführt. Dabei werden relevante Quellen aus den Verzeichnissen der vorherigen Treffer gesucht, was aus diesen Quellen ebenfalls wiederholt wird. Falls der Zugriff auf die Quellen über die oben genannten Plattformen (Google Scholar, Springer Link und ScienceDirect) nicht möglich ist, wird auf die lizenzierten e-Ressourcen der FHGR zurückgegriffen. Durch diese Rückwärtssuche können weitere relevante Treffer gefunden werden. Dieses Vorgehen wird auch «Schneeballverfahren» genannt (Döring & Bortz, 2016, S. 160).

Die Literaturrecherche erfolgt thematisch über mehrere Phasen. Allerdings muss bereits hier erwähnt werden, dass sich diese Phasen überschneiden werden, da bestimmte Literatur für mehrere Teile der Bachelorthesis relevant sein wird.

Phase 1: In einem ersten Schritt der Literaturarbeit werden die aktuellen Einsatzgebiete von KI im Recruiting analysiert. Dabei wird zuerst nachvollzogen, wie sich das Recruiting in den letzten Jahren bis zum Data-Driven-Recruiting, entwickelt hat, das gegenwärtig die modernste Möglichkeit der Personalrekrutierung darstellt (Jäger, 2018, S. 25). Danach werden die aktuellen Möglichkeiten von KI über die verschiedenen Rekrutierungsschritte vorgestellt. Aktuell wird KI bereits in unterschiedlichen Stufen des Recruitings verwendet (Teetz, 2021, S. 230). Es werden aktuelle Whitepapers zu KI im Recruiting analysiert, um nachzuvollziehen, wie die Empfehlungen aus der Praxis aussehen. Nach diesem Teil der Recherche wird angestrebt, die erste Unterfrage der Forschungsfrage beantworten zu können.

Phase 2: Im nächsten Schritt wird die Technologie hinter der KI genauer betrachtet. Dabei wird anhand von Fachliteratur ermittelt, wie durch Algorithmen ein Bias entstehen kann. Zuerst wird nachvollzogen, wie es allgemein zu einem Bias und danach spezifisch zu einem Genderbias im Recruiting kommen kann. Nach dieser Phase soll die zweite Unterfrage beantwortet werden.

Phase 3: In dieser Phase wird in der Literatur nach Massnahmen recherchiert, die eine diskriminierungsfreie KI fördern. Beispielsweise ist eine hohe Datenqualität in der Eingabe von KI-Systemen zentral für die Qualität der Ausgabedaten (Heim & Gerth, 2023, S. 121). Daher wird betrachtet, wie eine solche hohe Datenqualität erreicht werden kann. Zudem werden diverse weitere Faktoren recherchiert und analysiert. Des Weiteren sind technische Ansätze für eine diskriminierungsfreie KI vorhanden. Ein möglicher Ansatz bildet das neue Gebiet der erklärbaren KI (explainable AI oder XAI), das KI-Entscheide transparenter machen soll (Böhm et al., 2021, S. 198). Für diese Phase der Recherche wird die Literatur ab dem Jahr 2018 berücksichtigt, denn ab diesem Jahr erfolgte ein deutlicher Anstieg am Interesse der Explainable-Artificial-Intelligence, wodurch mehr wissenschaftliche Publikationen entstanden sind (Barredo Arrieta et al., 2020, S. 2). So wird sichergestellt, dass die Inhalte der Quellen nicht veraltet sind, was insbesondere hinsichtlich der technischeren Informationen und Anwendungen zur KI zentral ist.

In dieser Arbeit wird angestrebt, auf die geeignetsten Möglichkeiten spezifisch für KI-Tools im Rekrutierungsprozess einzugehen. Dies kann einerseits von der technischen Seite her erfolgen, um Handlungsempfehlungen für die Entwicklung von Tools abzuleiten, und andererseits aus der Sicht der User stattfinden (in diesem Fall der Recruiter), um für diese Empfehlungen zu geben, die sie im Bewerbungsprozess im Umgang mit KI-Tools beachten können. Die Handlungsempfehlungen werden daher konkret für diese beiden Zielgruppen erarbeitet. Nach der dritten Phase soll die Hauptforschungsfrage beantwortet werden.

2 Künstliche Intelligenz im Recruiting

Bei der KI geht es darum, mit computergestützten Verfahren Lösungen zu finden, für die sonst menschliche Intelligenz erforderlich wäre. Ballestrem et al. (2020) beschreiben die KI als «*Systeme, die intelligentes Verhalten zeigen, indem sie ihre Umgebung analysieren und – mit einem gewissen Grad an Autonomie – Massnahmen ergreifen, um bestimmte Ziele zu erreichen*» (S. 1). Im Kapitel 4 folgt eine grundlegende Einordnung und Erklärung der Technologie.

Bei einer Studie von index Research (2023, S. 7) gaben 12 % von 1128 befragten Unternehmen an, KI im Rekrutierungsprozess zu verwenden. Die überwiegende Mehrheit der Unternehmen setzt diese Technologie folglich noch nicht in der Personalabteilung und damit in der Rekrutierung ein. Bei Personalvermittlern wird KI bereits häufiger eingesetzt. Insgesamt 37,8 % der befragten Vermittler gaben an, KI einzusetzen. Am häufigsten kommt KI dabei für das Schreiben von Stellenanzeigen zum Einsatz. Bei einer Befragung von einzelnen HR-Fachpersonen gaben 70 % an, dass sie KI nutzen (Personio, 2023a, S. 3). Der Grund für diese hohe Diskrepanz zu den Werten der Unternehmensbefragungen ist, dass Recruiter solche Tools für ihre Aufgaben individuell nutzen, während in den Unternehmen noch keine definierten Workflows mit solchen Tools etabliert sind. In diesem Zusammenhang ist von Interesse, dass zahlreiche Bewerberinnen und Bewerber nicht wissen, dass KI im Recruiting eingesetzt wird. Laut einer Studie der Internationalen Hochschule IU (2022, S. 13) mit 1005 Befragten haben nur 6,3 % bisher bewusst Erfahrung damit gemacht. Ein Anteil von 88,5 % hat keine bewussten Erfahrungen gesammelt und 5,3 % wussten dies nicht.

Die Anwendung solcher Tools in der Rekrutierung umfasst in der Regel die Aufbereitung und Analyse umfangreicher Daten, um die Entscheidungsfindung zu erleichtern, den Bewerbungsprozess zu optimieren und die Kommunikation zwischen Unternehmen und Bewerbenden zu verbessern. Dabei werden während des gesamten Rekrutierungsprozesses Methoden der KI eingesetzt (Böhm et al., 2021, S. 195–198). Ein Hauptgrund dafür sind die gestiegenen Anforderungen an das Recruiting.

2.1 Anforderungen an das Recruiting

Auf dem heutigen Arbeitsmarkt stehen zahlreiche Unternehmen vor der Herausforderung, genügend geeignete Fachkräfte zu finden. Besonders in den IT-Berufen ist der Wettbewerb um die Talente besonders gross (Reindl & Krügl, 2023, S. 193). Da gute Fachkräfte um diesen Wettbewerb und damit um ihre gute Ausgangsposition auf dem

Arbeitsmarkt wissen, steigen ihre Ansprüche an den Bewerbungsprozess der Unternehmen (Böhm et al., 2021, S. 196). Ausserdem wird qualifiziertes Personal für Unternehmen immer wertvoller (Black & van Esch, 2020, S. 215). Seit etwa dem Jahr 2000 haben sich die Vorteile von Unternehmen von materiellen zu immateriellen Gütern wie dem Personal verlagert (Wilke & Bendel, 2022, S. 649). Durch die Digitalisierung hat sich der Bewerbungsprozess seit Ende der 1990er Jahre kontinuierlich weiterentwickelt. Für die Bewerbenden wurde es leichter, ihre Unterlagen schneller und effizienter zu versenden. Aus diesem Grund steigt die Zahl der Bewerbenden mit der Vereinfachung des Prozesses. Aufgrund dieser hohen Datenmenge ist es gegenwärtig oft nicht mehr möglich, die Aufgaben im Recruiting ohne die Hilfe von KI zu bewältigen (Black & van Esch, 2020, S. 217; Jares & Vogt, 2021, S. 75). Diese Entwicklung wird im folgenden Unterkapitel näher erläutert. Die Menge und Qualität der Daten ist ein zentraler Erfolgsfaktor für KI-Anwendungen. Da diese Werkzeuge datenzentriert arbeiten, hängt das Ergebnis der Ausgabedaten direkt von der Qualität der Eingabedaten ab. Eine Schwierigkeit besteht in der Abbildung bestimmter Informationen, z. B. Soft Skills, die sich nur schwer als Daten abbilden lassen. Ein weiteres Problem stellt das Blackbox-Modell dar, das die Nachvollziehbarkeit der Ausgabedaten verhindert (Böhm et al., 2021, S. 198). Darauf wird in Kapitel 4.2 eingegangen. Durch all diese gestiegenen Anforderungen hat sich das Recruiting in den letzten dreissig Jahren grundlegend gewandelt (Guenole & Feinzig, 2018, S. 6).

2.2 Entwicklung zum Data Driven Recruiting

Black und van Esch (2020, S. 217–218) zeigen die Entwicklung des digitalen Recruitings anhand von drei Evolutionsstufen auf.

1. **Digital Recruiting 1.0:** Das Digital Recruiting entstand Ende der 1990er Jahre durch die Digitalisierung und die Verbreitung des Internets. Dabei entstanden erste Plattformen wie monster.com. Durch die Entstehung von Netzwerkeffekten konnte der Bewerbungsprozess für beide Parteien effizienter gestaltet werden.
2. **Digital Recruiting 2.0:** Etwa zehn Jahre nach der flächendeckenden Verbreitung des Internets wurden die ersten grossen Social-Media-Plattformen bekannt. Im Jahr 2003 wurde LinkedIn gegründet, ein Jahr später folgte Facebook. Durch diese Plattformen wuchsen die Netzwerkeffekte exponentiell und potenzielle Arbeitnehmende konnten zunehmend erreicht werden. Durch die sozialen Medien und die vereinfachte Bewerbung stiegen die Ansprüche der Bewerberinnen und Bewerber an einen effizienten Bewerbungsprozess (Guenole & Feinzig, 2018). In dieser Phase entstanden erste zentralisierte Jobbörsen wie indeed.com.

3. **Digital Recruiting 3.0:** In den Phasen 1.0 und 2.0 wurde es für Talente zunehmend unkomplizierter, Bewerbungen zu versenden. Dadurch stieg die Anzahl der Bewerbungsdossiers stetig an, was 2015 zur Phase 3.0 führte, dem Data-Driven-Recruiting. Die KI gilt als zentrales Element in dieser Entwicklung, da die Technologie ab dieser Phase eingesetzt wurde, um die hohe Menge an Bewerbungen bearbeiten zu können (Z. Chen, 2023, S. 139). Derzeit besteht diese Phase weiterhin fort.

2.3 Aktueller Einsatz von Künstlicher Intelligenz

Black und van Esch (2020, S. 218) sowie Z. Chen (2023) zeigen auf, dass KI-Tools bereits über alle Stufen des Recruitings eingesetzt werden. Black und van Esch unterteilen den Prozess dabei in vier Stufen: Kontaktaufnahme, Screening, Bewertung und Koordinierung. Z. Chen unterteilt den Prozess genauer und definiert sechs Phasen: Stellenanzeige, Stellensuche, Bewerbung, Screening, Bewertung und Koordination (Abbildung 1). Für diese Arbeit werden die Prozessschritte von Z. Chen verwendet, da diese trennschärfer sind und somit eine spätere Handlungsempfehlung präziser möglich ist.

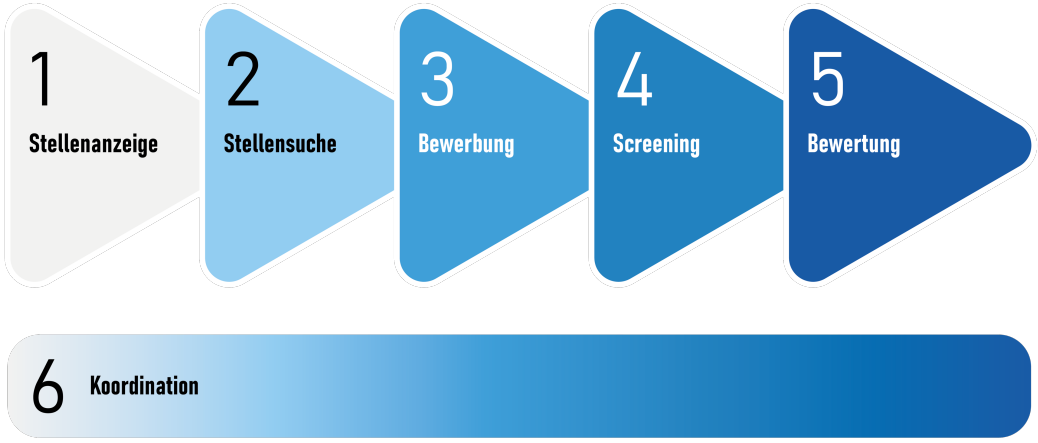

Abbildung 1: Phasen des Recruitings, (Eigene Darstellung in Anlehnung an Z. Chen, 2023)

Phase 1, Stellenanzeige: Stellenanzeigen können mithilfe von KI erstellt werden. So kann die Stellenbeschreibung automatisch generiert oder verbessert werden. Darüber hinaus kann KI analysieren, welche externen Kanäle für die Stellenausschreibung am effektivsten sind und wie Kandidatinnen und Kandidaten vorzugsweise angesprochen werden. Die Anzeige kann für jede Person individualisiert werden, wodurch der Inhalt ansprechender auf die Zielperson zugeschnitten werden kann (Z. Chen, 2023, S. 140). Die erstellten Inserate können durch KI zudem einen grösseren Bewerbenden-Pool erreichen. Die Anzahl der passiven Bewerbenden ist dreimal so hoch wie die der aktiven. Eine Vielzahl davon ist offen für einen neuen Job, auch wenn sie nicht aktiv danach sucht. Zudem finden sich deutlich mehr Informationen zu den Kandidatinnen und Kandidaten

als früher. Diese Talente können mithilfe der KI-Technologie erreicht werden (Black & van Esch, 2020, S. 219). Durch Programmatic-Job-Advertising wird dabei das Stelleninserat mithilfe von KI zum richtigen Zeitpunkt an die richtige Person ausgespielt (Hasenbein, 2023, S. 90). Mithilfe von KI kann in dieser Phase bereits berechnet werden, wie lange es durchschnittlich dauert, bis eine solche Stelle besetzt ist (Guenole & Feinzig, 2018, S. 12). Zudem ist die Akzeptanz von KI bei den Bewerbenden in der Phase der Stellenausschreibung am höchsten. Je weiter fortgeschritten der Bewerbungsprozess ist, desto geringer ist die Akzeptanz für solche Tools. Insgesamt 69,8 % der befragten Bewerbenden sehen KI in dieser Phase als positiv an (IU Internationale Hochschule, 2022, S. 13).

Phase 2, Stellensuche: Bei der Stellensuche kann KI die Suchergebnisse (beispielsweise auf LinkedIn oder Google) präziser gestalten und somit passendere Stellen für Bewerbende vorschlagen. Diese können ausserdem einen Erstkontakt mit einem Chatbot aufnehmen, welcher grundlegende Fragen zu einer Stelle beantworten kann (Z. Chen, 2023, S. 140). Wie bei den Stellenanzeigen ist die Akzeptanz für Chatbots zu Beginn des Prozesses mit 58,6 % eher hoch (IU Internationale Hochschule, 2022, S. 13).

Phase 3, Bewerbung: Beim Schritt der Bewerbung kann durch die Hilfe von Algorithmen der Aufwand für Recruiter und Bewerbende reduziert werden. Einerseits können Informationen von Bewerbenden direkt in einem System hinterlegt und müssen somit nicht manuell eingegeben werden, andererseits können hochgeladene Dokumente wie ein CV automatisch gelesen und strukturiert abgelegt werden (Z. Chen, 2023, S. 140).

Phase 4, Screening: Ein automatisiertes Screening durch KI hilft der HR-Abteilung dabei, Kandidatinnen und Kandidaten effizienter zu bewerten und vollständig ungeeignete Bewerbungen direkt auszusortieren. Dabei können gamifizierte Tests helfen, die Performance und das Wissen einer bewerbenden Person schnell einzuschätzen (Z. Chen, 2023, S. 140–141). Bewerbungen können dabei automatisch gescreent werden und bis zu 75 % der unbrauchbaren Bewerbungen werden direkt aussortiert (Kambur & Yildirim, 2022, S. 93). Vor allem die Time-to-Hire wird in dieser Phase durch die KI verkürzt. Dadurch entsteht ein Wettbewerbsvorteil in Märkten mit hoher Fluktuation (Black & van Esch, 2020, S. 220).

Phase 5, Bewertung: In diesem Prozessschritt können die Anforderungen an die Stelle automatisiert mit den Angaben in der Bewerbung verglichen werden, wodurch ein sogenannter «Resume-Score» erstellt werden kann. Darüber hinaus können aufgezeichnete Videointerviews durch KI analysiert und ausgewertet werden (Z. Chen, 2023, S. 140). Hierbei wird beispielsweise anhand der Stimme, Körpersprache oder Wortwahl

ausgewertet, ob die Person zur Stelle passt oder nicht (Jares & Vogt, 2021, S. 76; Kambur & Yildirim, 2022, S. 94). Bei der Videoanalyse durch KI ist die Akzeptanz der Bewerbenden bereits deutlich geringer als zu Beginn des Prozesses. Ein Anteil von 61,8 % bewertet diesen Einsatz negativ (IU Internationale Hochschule, 2022, S. 13).

Phase 6: Koordination: Die Koordination gilt hier nicht als abschliessende Phase, sondern findet (wie in Abbildung 1 dargestellt) parallel zu allen Phasen statt. Dabei ist das Ziel, den Bewerbenden über den gesamten Prozess hinweg eine angenehme Journey zu bieten und dadurch auch für ungeeignete Kandidatinnen und Kandidaten ein positives Image sicherzustellen (Black & van Esch, 2020, S. 221). Insbesondere abgelehnte Bewerbende sollen eine positive Erfahrung aus dem Bewerbungsprozess mitnehmen, da sie möglicherweise für eine zukünftige Stelle geeignet sein könnten (Z. Chen, 2023, S. 141).

Zusammengefasst kann gesagt werden, dass KI den Personalabteilungen dabei hilft, möglichst zahlreiche und geeignete Bewerbungen zu erhalten und diese effizienter und für die Bewerbenden angenehmer zu bearbeiten. Dabei handelt es sich mehrheitlich um unterstützende KI-Anwendungen. Die Recruiter werden in ihren Aufgaben nicht vollständig ersetzt (Böhm et al., 2021, S. 214).

2.4 Gründe für den Einsatz von Künstlicher Intelligenz im Recruiting

Einer der grössten Vorteile im Recruiting ist die Effizienzsteigerung durch KI-Tools (Black & van Esch, 2020, S. 222). Durch eine Automatisierung einzelner Arbeitsschritte entsteht dabei eine grosse Kostenreduktion (Onlyfy, 2023, S. 5). Diese findet derzeit vor allem in den Phasen 1 (Stellenausschreibung) und 4 (Screening) statt (Wilke & Bendel, 2022, S. 655). Insgesamt 93 % der Personalverantwortlichen, die KI in ihrem Unternehmen einsetzen, glauben daran, dass sie dadurch Kosten einsparen (Personio, 2023a, S. 4). Aktuell geben 64 % der HR-Fachpersonen an, dass sie vorwiegend für operative Verwaltungsaufgaben verantwortlich sind (Personio, 2023a, S. 16). Durch die Zeiteinsparung durch KI können sie sich somit beispielsweise auf strategische oder zwischenmenschliche Arbeiten konzentrieren, bei denen die KI noch weniger hilfreich ist (Guenole & Feinzig, 2018, S. 7). Durch KI-Tools ist es auch möglich, eine deutlich grössere Anzahl an Bewerbungen zu bearbeiten, als es durch Menschen möglich wäre, z. B. durch aktives Sourcing und die Ansprache passiver Kandidatinnen und Kandidaten (Wilke & Bendel, 2022, S. 656). Zudem kann eine höhere Anzahl an Datenpunkten pro Bewerbung berücksichtigt werden (Onlyfy, 2023, S. 5). Ein weiterer Grund für den Einsatz von KI ist die Verbesserung der Customer-Journey. Wie bereits im Kapitel 3.3 erläutert wurde, helfen

solche Tools dabei, den Bewerbenden einen personalisierten und angenehmen Bewerbungsprozess zu bieten (Guenole & Feinzig, 2018, S. 7). Auch in der Recruiting-Branche werden unterschiedliche Gründe und Einsatzszenarien für die KI genannt. Diese werden vermehrt in der Form von Whitepapers veröffentlicht.

2.5 Empfehlungen für die HR-Abteilung aus Whitepapers

Für diesen Abschnitt wurden drei Whitepapers sowie ein Interview zu KI im Recruiting analysiert, um zu sehen, wie die Empfehlungen aus der Praxis aussehen. Die Papers stammen von den Firmen Onlyfy (Bewerbungsmanager von Xing), Personio (HR-Softwarefirma), und employ (Softwarefirma und Anbieter des KI-Tools Jobvite). Das Interview stammt aus dem Magazin «Recruiting now» (Branchenmagazin von Personal Schweiz) und wurde mit Roger Basler de Roca geführt.

In den Whitepapers werden jeweils unterschiedliche Anwendungsmöglichkeiten aufgezeigt, wobei sich manche davon überschneiden. Dabei ist die Möglichkeit für das automatische Generieren von Stellenanzeigen hervorzuheben. Diese wird in jedem Paper als passendes Einsatzgebiet erwähnt (Basler de Roca, 2023; Employ & JOBVITE, 2023; Onlyfy, 2023; Personio, 2023a). Auch die Möglichkeiten des Active Sourcings, der Kommunikation mit Bewerbern und des Job-Profile-Matchings werden in drei der vier Paper erwähnt. Als weiteres mögliches Einsatzszenario wird die automatisierte Interviewplanung oder der Einsatz von Chatbots genannt.

In allen Dokumenten wird erwähnt, dass der Einsatz des Menschen nicht verlorengehen darf und die KI als Unterstützung angesehen werden soll – auch vor dem Hintergrund, dass 80 % der Bewerberinnen und Bewerber spätestens im Vorstellungsgespräch eine reale Person sehen möchten. Insgesamt 37 % der Befragten lehnen die Kommunikation mit einem Chatbot ab, egal in welcher Bewerbungsphase. Dabei wurden rund 1000 Erwerbstätige zwischen 18 und 65 Jahren befragt (Onlyfy, 2023, S. 6).

Ebenfalls wird in allen Berichten betont, dass trotz der Zeitersparnis durch KI eine Qualitätssicherung durch eine HR-Fachkraft notwendig ist. Darüber hinaus sollten die Anwendungen und deren Ergebnisse in regelmässigen Abständen (z. B. in einer monatlichen Feedbackschleife) im Team diskutiert und kritisch hinterfragt werden. Gegebenenfalls sollte der Einsatz solcher Tools angepasst werden.

Die Thematik der Diskriminierung durch einen Bias in KI-Tools wird in jedem der Whitepaper aufgegriffen. Die Firma Onlyfy (2023, S. 5) warnt in ihrem Whitepaper, dass die Gefahr der Diskriminierung auch bei KI besteht. Konkrete Handlungsanweisungen für Personalverantwortliche, wie damit umzugehen ist, fehlen jedoch. Ähnlich sieht es in der

Schweizer Branchenzeitschrift «Recruiting Now» aus. Basler de Roca (2023, S. 6) warnt im Interview vor einem möglichen Bias durch KI und die damit einhergehenden ethischen Probleme. Auch demnach ist es schwierig, als Recruiter zu wissen, wie ein verantwortungsbewusster Umgang mit der KI sichergestellt werden kann. Im Whitepaper der Firma Personio (2023a) wird auf die Gefahr einer Voreingenommenheit (Bias) hingewiesen, aber nicht am Beispiel der Geschlechterdiskriminierung. Als Handlungsempfehlung gegen ein Bias wird mitgegeben, die Tools nicht mit bestehenden Daten zu befüllen, sondern mit wünschenswerten idealen Indikatoren. In der Praxis könnte dies jedoch schwierig umzusetzen sein, da Datenmengen in grosser Menge und guter Qualität benötigt werden (Heim & Gerth, 2023, S. 121–122). Auch die Firma Employ äussert Bedenken hinsichtlich der Fairness von KI-Algorithmen im Recruiting (2023, S. 16–17). Sie weist auf die Notwendigkeit hin, die Technologie und die Ergebnisse kritisch zu hinterfragen. Die KI könne auch dabei helfen, mehr Vielfalt im Recruiting zu erreichen.

Wie an den Beispielen dieser verschiedenen Whitepapers ersichtlich ist, wird das Problem erkannt und angesprochen. Die Handlungsempfehlungen hingegen sind nicht konkret genug, damit das HR-Personal damit einen möglichen Bias verhindern kann. Nebst der kritischen Perspektive aus der Recruiting-Branche tragen auch negative öffentliche Fälle aus der Praxis dazu bei, dass das Image von KI nicht nur positiv ist.

2.6 Beispiele von aufgetretenem Genderbias

Amazon: Der Online-Versandhändler Amazon hatte im Jahr 2015 insgesamt 110’000 Stellen zu besetzen. Da dies eine deutlich höhere Zahl war als noch im Jahr 2014 (77’000 Stellen) entwickelte das Unternehmen ein KI-Tool, das es dabei unterstützen sollte. Für die Entwicklung des Tools wurden grosse Datenmengen des Unternehmens aus den letzten zehn Jahren verwendet (Black & van Esch, 2020, S. 223). Das Unternehmen musste das KI-Rekrutierungstool jedoch wieder einstellen, da es Frauen für bestimmte Berufe, besonders in technischen Bereichen, systematisch schlechter einstufte als Männer. Diese Voreingenommenheit war darauf zurückzuführen, dass die Musterdaten für «passende» Bewerber überwiegend von Männern stammten. So wurde der Begriff «woman» oder Schulen, die nur von Frauen besucht werden, schlechter eingestuft (Orwat, 2019, S. 34). Daraus schloss der Algorithmus, dass Frauen für bestimmte Berufe weniger geeignet sind (Lavanchy, 2018).

Jobplattformen: In einer Studie (L. Chen et al., 2018) wurde aufgezeigt, dass auch auf Plattformen für die Stellensuche ein Bias gegenüber Frauen auftreten kann. Auf den Plattformen Indeed, Monster und CareerBuilder wurden für 35 Berufsbezeichnungen die

Daten von rund 855’000 Kandidatinnen- und Kandidatenprofile verwendet, um zu analysieren, wie diese vom System eingestuft werden. In 12 der 35 Jobbezeichnungen erfolgte eine signifikante Bevorzugung männlicher Profile gegenüber weiblichen Profilen. Auch auf den Plattformen TaskRabbit und Fiverr wurde ein diskriminierender Bias gegen Frauen nachgewiesen. Auf diesen Plattformen werden kleinere Arbeiten für Freiberuflerinnen und Freiberufler angeboten. Dabei wurde einerseits entdeckt, dass Frauen auf ihren Profilen 10 % weniger Bewertungen erhalten als Männer, und andererseits, dass sie in der Rangfolge der Suche tiefer unten erscheinen – selbst bei vergleichbarer Qualifikation (Orwat, 2019, S. 35).

Social Media: Im Jahr 2018 wurde Facebook von der Bürgerrechtsorganisation American Civil Liberties Union angeklagt, da das Unternehmen bestimmte Stelleninserate ausschliesslich an männliche Personen ausgespielt hatte. Das Merkmal «Geschlecht» ist bei der Ausschreibung von Stellenanzeigen zwar geschützt, doch der Fehler entstand bereits vorher. Unternehmen konnten Facebook Datensätze über ihre Mitarbeitenden zur Verfügung stellen, um dadurch die Jobinserate gezielter ausspielen zu können. Dabei erstellt Facebook eine sogenannte «Lookalike-Audience», die sich aus den bisherigen demografischen Daten abgeleitet hatte. Wie beim Beispiel von Amazon waren die Mitarbeitenden bei bestimmten Unternehmen überwiegend männlich, wodurch sich auf Facebook ein geschlechtsspezifischer Bias manifestierte (Orwat, 2019, S. 40). Welche Arten von Bias in diesen Fällen aufgetreten sind, wird in Kapitel 4.3.1 erläutert.

Darüber hinaus finden sich weitere Fälle von aufgetretenem Geschlechterbias. Beispielsweise wurde in einem Experiment herausgefunden, dass tiefer bezahlte Stellen vorgeschlagen werden, wenn das Geschlecht in den Google-Ads-Einstellungen auf «weiblich» gestellt wird (Datta et al., 2015, S. 3). Auch aufgrund solcher Cases sehen auch Expertinnen und Experten den Umgang mit KI in der Rekrutierung kritisch an.

2.7 Kritik an KI-Tools im Recruiting

Auch wenn KI-Tools, wie in Kapitel 3.4 erläutert, für Unternehmen Vorteile mit sich bringen, treten insbesondere aus ethischer Sicht Bedenken auf. Rekrutierungs-Expertin Natalie Gyöngyösi (2022) kritisiert den Einsatz von KI im Recruiting, da die Technologie nicht selbst urteile, sondern Fehler und falsche Rollenbilder reproduziere:

> *Die AI urteilt nicht. Sondern sie kopiert sinnlos sowohl wohlwollendes Verhalten wie auch das umgekehrte. Wenn ihr von einem menschlichen Rollenmodell rassistisches, frauenverachtendes oder geschlechterdiskriminierendes Futter vorgeworfen wird, frisst sie dieses genauso*

> *frischfröhlich wie alles andere. […] Das Resultat ist eine AI mit lauter positiven wie negativen Stereotypen «im Kopf» – ein Abbild der Mitglieder unserer Gesellschaft. (S. 26–27)*

Auch der Arbeits- und Organisationspsychologe Gian-Rico Bardy (2022) warnt vor den Gefahren. Er erläutert, dass bei einem fehlenden Verständnis für die Funktionsweise von KI-Algorithmen das Risiko in Kauf genommen wird, dass die KI aufgrund mangelnder Datenlage diskriminierende Entscheidungen treffen kann. Wie es zu diesem Unverständnis der KI-Technologie kommen kann, wird in Kapitel 4.2 aufgezeigt.

2.7.1 Einstellung der Bewerbenden

Wie bei den Expertinnen und Experten ist die Einstellung der Bewerberinnen und Bewerber eher negativ. Bei einer Studie der Internationalen Hochschule UI (2022) gaben von 1005 befragten Personen zwischen 16 und 65 Jahren 65,2 % an, negative Emotionen gegenüber KI im Recruiting zu haben. Nur 22,7 % haben positive Emotionen und der übrige Teil ist unentschlossen oder hat indifferente Emotionen. Nahezu die Hälfte der Befragten (47,2 %) sieht dabei einen Nachteil in der «Anfälligkeit von KI für Verzerrungen und Stereotypen». Dieser Punkt wird nach «Unpersönlichkeit» und «Fehler in der Programmierung der KI» als drittgrösster Nachteil betrachtet. Dies zeigt auf, dass diese Thematik auch bei den Bewerbenden präsent ist und ein Handlungsbedarf auf der Seite der Entwicklung und Rekrutierung besteht. Zudem gaben 43 % der Befragten an, zu glauben, dass sich der Bewerbungsverlauf für sie durch KI verschlechtert. Nur 18,6 % denken, dass sich der Prozess dadurch für sie verbessert. Der übrige Teil sieht den Einfluss der KI als neutral an. Grundsätzlich wird von den Bewerbenden die Transparenz der Datenverarbeitung über alle Schritte im Bewerbungsprozess hinweg gefordert. Für sie ist dabei die Gleichstellung und Transparenz von hoher Bedeutung (Böhm et al., 2021, S. 201).

2.7.2 Einstellung der Recruiter

Die Meinungen von Recruitern gegenüber dem Einsatz von KI bei ihrer Arbeit sind ambivalent. Sie erkennen an, dass die KI die tägliche Arbeit effizienter gestalten kann und somit mehr Zeit für zwischenmenschliche und strategische Arbeiten bleibt. Insgesamt 70 % der befragten HR-Fachpersonen (Personio, 2023b, S. 23) gaben an, das KI das Potential bietet, mehr Zeit für sonstige Arbeit einzuräumen. Noch mehr (73 %) sind der Meinung, die Technologie werde benötigt, um mit der Arbeitsbelastung mithalten zu können. Bei der Befragung von Employ (2023, S. 15) gaben zudem 42 % der Fachpersonen an, dass sie glauben, durch KI strategischer in ihrer täglichen Arbeit vorgehen zu können.

Neben den positiven Aspekten sind sich die Personalverantwortlichen aber auch der negativen Aspekte bewusst. Ein Anteil von 67 % der Befragten ist generell besorgt über die Auswirkungen von KI auf Arbeitsplätze und Qualifikationsanforderungen und 57 % sind sogar besorgt um einen persönlichen Jobverlust (Personio, 2023b, S. 23). Des Weiteren besteht die Gefahr, das durch KI-Tools die jahrelange Erfahrung menschlicher Recruiter ausgeblendet wird und beispielsweise das Gefühl, ob eine Person zum Unternehmen passt, nicht mehr berücksichtigt wird (Christen et al., 2020, S. 160). Ebenso wie für die Bewerbenden ist auch für die Recruiter von Bedeutung, dass trotz KI weiterhin ein persönlicher Kontakt zustande kommt (Thalmann et al., 2022, S. 1).

In Kapitel 2 wurde der aktuelle Einsatz von KI im Recruiting analysiert und aufgezeigt, dass KI trotz Effizienzsteigerungen und Kosteneinsparungen Herausforderungen wie Geschlechterbias und mangelnde Transparenz mit sich bringt. Es wurden die Entwicklungsstufen des digitalen Recruitings und die zentrale Rolle von KI im heutigen Data-Driven-Recruiting beschrieben, wobei auch die negativen Einstellungen der Bewerbenden und Recruiter gegenüber KI-Tools beleuchtet wurden. Im folgenden Kapitel wird die Technologie der KI detailliert analysiert, mit besonderem Fokus auf Machine-Learning und die Problematik des Blackbox-Modells.

3 Funktionsweise Künstlicher Intelligenz

Auch wenn die KI durch Tools wie ChatGPT in den letzten Jahren stark an Bekanntheit gewonnen hat, reichen die ersten Ideen einer KI bereits Jahrzehnte zurück. In ihrem Artikel «A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence» (1955) beschreiben McCarthy et al. bereits vor nahezu siebzig Jahren Ideen und Möglichkeiten einer solchen Technologie. Seither wurde der Begriff der KI zunehmend bekannter und in den letzten Jahren inflationär verwendet. Besonders grosse Fortschritte wurden in den letzten Jahren durch Deep Learning (tiefes Lernen) erzielt (Böhm et al., 2021, S. 197). Der Begriff Deep Learning wird in Kapitel 4.1.3 erklärt. Durch KI werden automatisierte Prozesse für eine Lösungserarbeitung definiert, die auf Algorithmen basieren. Letztere sind Anweisungen für den Computer, die sequenziell und in hohem Tempo ausgeführt werden. Computer funktionieren nach dem EVA-Prinzip, bei dem zuerst Daten eingegeben (E), dann verarbeitet (V) und zuletzt ausgegeben (A) werden. Algorithmen und dadurch die KI kommen dabei in der Verarbeitung der Daten zum Einsatz (Krebs & Hagenweiler, 2022, S. 8). Sie übersetzen dabei mithilfe mathematischer Operationen Anweisungen in Computercode um und sind für zahlreiche moderne Errungenschaften technischer Maschinen verantwortlich (Fry, 2019, S. 20).

3.1 Einteilung der KI

Um in einem späteren Schritt die Entstehung eines Bias besser verstehen zu können, wird in einem ersten Schritt die Technologie der KI erklärt und eingeordnet. Dabei sind verschiedene Begriffe zu verstehen. Auch wenn die Begriffe «Künstliche Intelligenz», «Machine-Learning» (maschinelles Lernen) und «Deep Learning» oft synonym verwendet werden, sind sie nicht dasselbe (Wuttke, 2023).

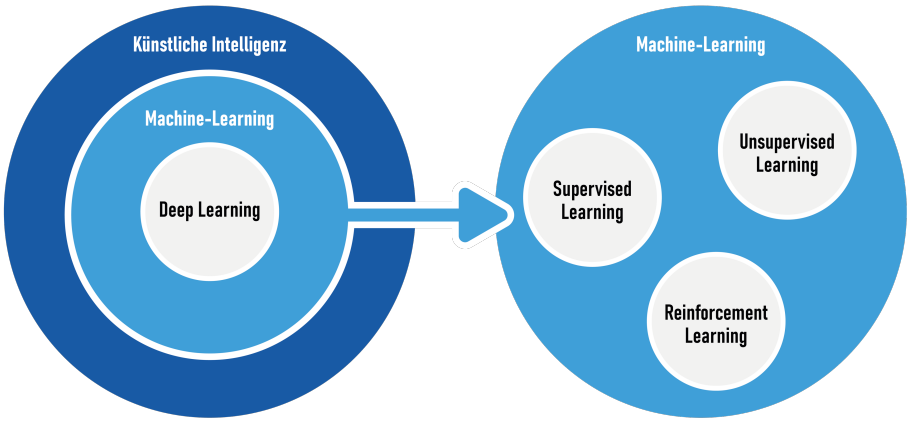

Abbildung 2: Einordnung Künstliche Intelligenz, Machine-Learning und Deep Learning, (Eigene Darstellung angelehnt an Wuttke, 2023)

Wie in Abbildung 2 erkennbar ist, ist Machine-Learning ein Teilbereich der KI und Deep Learning ein Teilbereich des Machine-Learnings. Letzeres basiert dabei auf der Technologie der Neural Networks (neuronale Netzwerke) und kann in drei Bereiche unterteilt werden: Supervised Learning, Unsupervised Learning und Reinforcement-Learning (Christen et al., 2020, S. 86). Die Begriffe werden in den folgenden Unterkapiteln erläutert.

3.1.1 Neural Networks

Das Machine-Learning basiert auf der Technologie der Neural Networks (neuronale Netzwerke). Diese Netzwerke bestehen aus künstlichen Neuronen, die zu Knotenpunkten geformt werden und dadurch der Struktur der Neuronen im menschlichen Gehirn nachempfunden sind (Krebs & Hagenweiler, 2022, S. 14). Die Intention besteht darin, ein mathematisches Modell des Gehirns abzubilden (Suzuki, 2011, S. 3).

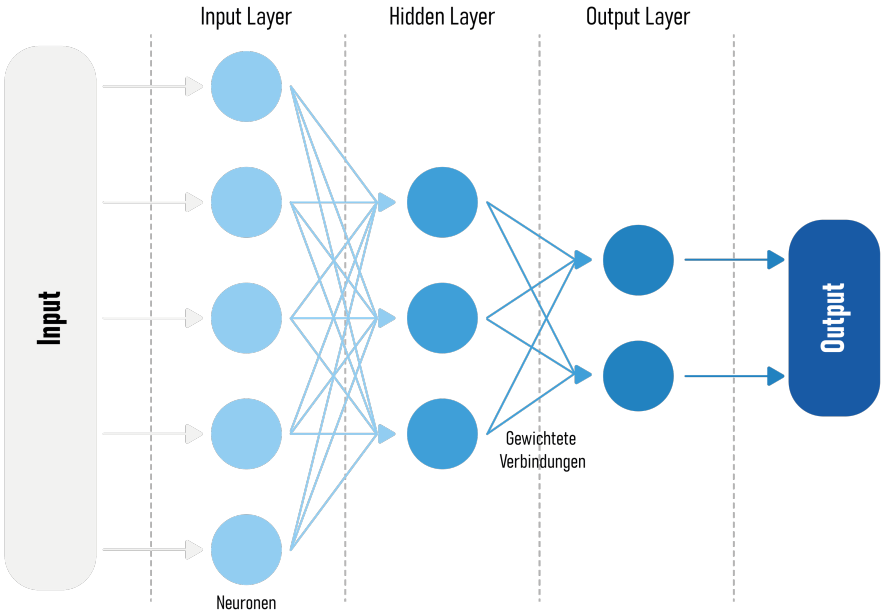

Abbildung 3: Neural Networks, (Eigene Darstellung angeleht an Dike et al., 2018, S.322; Suzuki, 2011, S. 6)

Wie in Abbildung 3 erkennbar ist, lassen sich die Neural Networks in drei Ebenen unterteilen. Der «Input-Layer» repräsentiert die Verbindung zur Aussenwelt. Im Rahmen dessen erfolgt eine Sammlung von Inputs, die verarbeitet werden können. Die Verarbeitung erfolgt im «Hidden Layer», der sämtliche Neuronen umfasst. Die Neuronen sind dabei durch Gewichtungen miteinander verbunden, wobei die Anzahl der Verbindungen variiert. Die Gewichtungen in den Verbindungen können durch verschiedene Methoden angepasst werden, was zu unterschiedlichen Ergebnissen führt. Die Gewichte des Netzwerks werden durch die vorgenommenen Anpassungen modifiziert, wodurch das

Netzwerk lernt (Suzuki, 2011, S. 3). Wie diese Anpassungen vorgenommen werden, wird in den nächsten Unterkapiteln erläutert. Zum Ende folgt der «Output-Layer», in dem die verarbeiteten Daten in einer geeigneten Form wieder an die Aussenwelt abgegeben werden (Dike et al., 2018, S. 322).

Ein wesentlicher Vorteil solcher Neural Networks besteht in der Möglichkeit, dass das System auf die Aussenwelt zugreift und daraus lernt. Dies ist insbesondere von Vorteil, wenn der Kontext eine hohe Komplexität aufweist und folglich die Anwendung von Vereinfachungen durch den Menschen erschwert wird. Die Technologie der Neural Networks ist dazu in der Lage, eine Vielzahl von Aufgaben der Datenverarbeitung zu erfüllen. Mögliche Anwendungen sind beispielsweise die Classification (Klassifizierung), das Filtering (Filtern) oder das Decision-Making (Entscheiden) aufgrund von Inputdaten (Suzuki, 2011, S. 14)

3.1.2 Machine-Learning

Das Machine-Learning bildet einen Teilbereich der KI. Dabei erlernt das KI-System auf Basis bestehender Daten, Muster zu erkennen, um daraus selbst Lösungen zu entwickeln. Zum Entwickeln eines solchen Systems wird ein Trainingsdatensatz verwendet, an dem Muster und Zusammenhänge der Daten erkannt werden sollen. Das Erlernte wird danach anhand von Test- und Validierungsdaten geprüft und korrigiert. Dadurch kann die Leistung des Algorithmus so lange optimiert werden, bis die Ergebnisse daraus zufriedenstellend sind. Die drei Arten des Machine-Learnings werden in den folgenden Abschnitten erklärt.

3.1.2.1 Supervised Learning

Beim Supervised Learning (überwachtes Lernen) wird der Machine-Learning-Algorithmus anhand von sogenannten Benchmark-Daten trainiert (Abbildung 4). Diese Daten werden dabei als korrekte Lösung eines Outputs angesehen. Folglich ist bereits vor dem Supervised Learning klar, wie die Zieldaten aus dem KI-System aussehen sollen. Das Training wird dabei von einem Menschen überwacht, wie der Name verrät. Die Herausforderung besteht dabei in der Identifizierung der richtigen Merkmale (Benchmark-Daten). Diese Aufgabe wird von Expertinnen und Experten des Fachgebietes erledigt und nennt sich Feature-Engineering (Christen et al., 2020, S. 86). Die Eingabedaten werden klassifiziert und gelabelt, sodass sie mit den Ausgabedaten verglichen werden und der Lernerfolg der KI bewertet werden kann (Krebs & Hagenweiler, 2022, S. 12). Dieser Schritt des Lernens wird mehrmals wiederholt. In diesen Iterationen werden die Gewichte des Neural Networks (vgl. Abbildung 3) so angepasst, dass sich der Eingabewert (−)

zunehmend dem Zielwert (+) nähert und die geringste Differenz erreicht wird (Dike et al., 2018, S. 323).

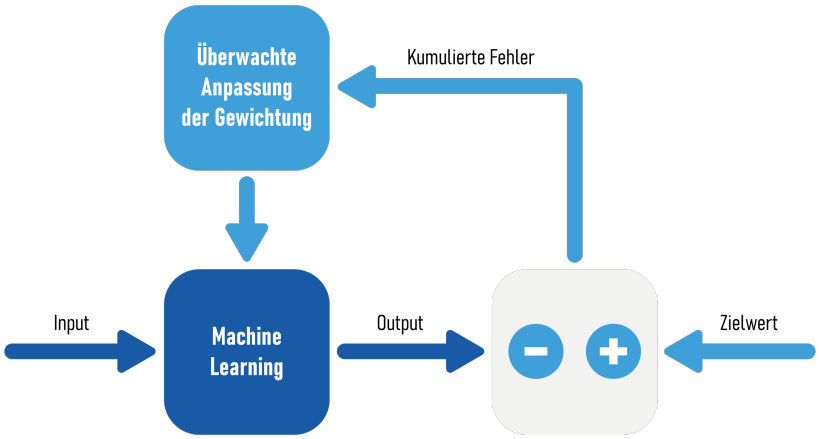

Abbildung 4: Supervised Learning, (Eigene Darstellung angelehnt an Dike et al., 2018, S.323)

3.1.2.2 Unsupervised Learning

Im Gegensatz zum Supervised Learning sind beim Unsupervised Learning (unüberwachtes Lernen) die korrekten Ausgabedaten noch nicht bekannt. Die Eingabedaten werden deshalb nicht gelabelt und mit den Ausgabedaten direkt verglichen (Abbildung 5). Bei diesem Typ des Machine-Learnings werden grössere Datenmengen verarbeitet als in den anderen Verfahren (Böhm et al., 2021, S. 198). Durch diese grösseren Datenmengen und einer schnelleren Verarbeitungs- und Lernzeit bietet dieses Machine-Learning-Verfahren neue Möglichkeiten im Bereich der KI (Dike et al., 2018, S. 322). Aus den Eingabedaten werden dabei vom Machine-Learning-Algorithmus automatisch Zusammenhänge und Abhängigkeiten herausgesucht, ohne dass ein Mensch darauf Einfluss nimmt (Jörgens et al., 2020, S. 142). Dies findet beispielsweise durch sogenanntes Clustering statt, bei dem bestimmte Datensätze als zusammenhängend wahrgenommen und gruppiert werden (Christen et al., 2020, S. 87). Durch dieses Clustering kann ein Bias entstehen, da Datensätze zusammen assoziiert werden, die zu einem diskriminierenden Ergebnis wie bei Amazon führen (vgl. Kapitel 3.6).

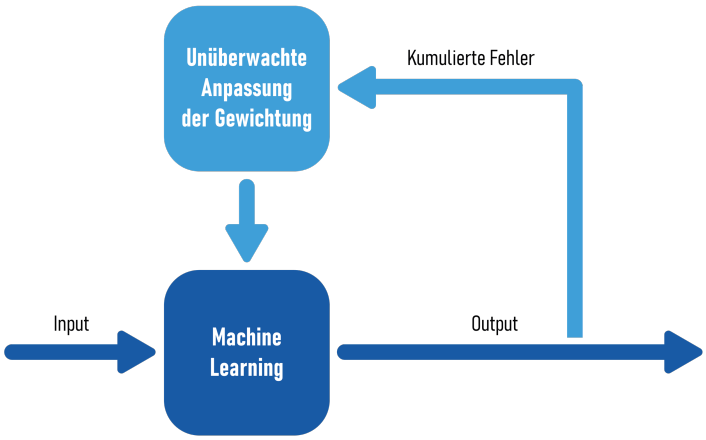

Abbildung 5: Unsupervised Learning, (Eigene Darstellung angelehnt an Dike et al., 2018, S. 324)

3.1.2.3 Reinforcement Learning

Das Reinforcement-Learning (Abbildung 6) basiert auf einer Verhaltenspsychologie, die auch beim Menschen beobachtet werden kann. Bei einem positiven Resultat wird eine Belohnung ausgesprochen, während bei einem negativen Resultat eine Bestrafung erfolgt. Wie beim Unsupervised Learning ist auch hier nicht bekannt, wie die optimalen Ausgabedaten aussehen sollen. Das Ziel dieser Lernmethode ist es, dass das System eine optimale Lösung aus den Eingabedaten auswählt. Wie bei den anderen Verfahren erfolgt der Lernprozess dabei iterativ. Nach jedem Schritt wird das System unterschiedlich stark belohnt oder bestraft. Das Ziel des Systems ist es dabei, eine möglichst hohe Belohnung und dadurch ein optimales Ergebnis zu erzielen (Christen et al., 2020, S. 87). Der Prozess erfolgt nach einem Trial-Error-Prinzip so lang, bis die Ausgabe zufriedenstellend ist (Krebs & Hagenweiler, 2022, S. 13). Die Herausforderung besteht darin, ein Gleichgewicht zwischen der Nutzung der aktuell verfügbaren Daten und der Entdeckung neuer Informationen herzustellen (Dike et al., 2018, S. 323). Wie an der Abbildung 6 zu erkennen ist, ist das Reinforcement-Learning dem Unsupervised Learning ähnlich, nur dass das Signal bei der Anpassung der Gewichtung nicht innerhalb des Systems selbst erfolgt, sondern von einer Person durchgeführt wird.

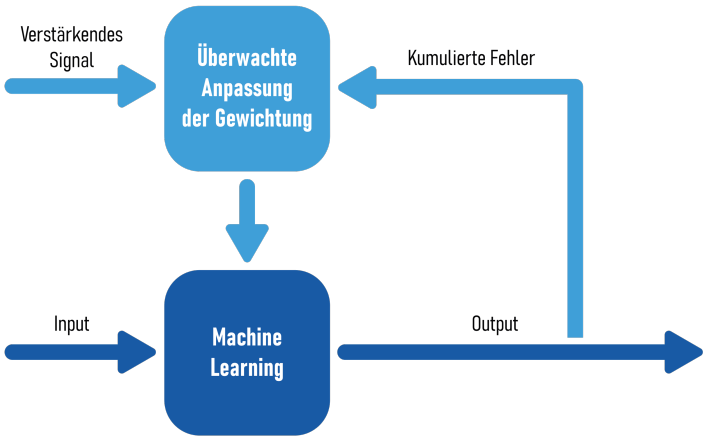

Abbildung 6: Reinforcement Learning, (Eigene Darstellung angelehnt an Dike et al, 2018, S.323)

3.1.3 Deep Learning

Deep Learning kann als Unterkategorie des Machine-Learnings betrachtet werden. Dabei basiert es ebenfalls auf Neural Networks, die beim Deep Learning jedoch eine höhere Anzahl an Schichten enthalten als beim Machine-Learning (Christen et al., 2020, S. 86). Durch diese mehrschichtigen Netzwerke kann beim Deep Learning eine deutlich grössere Anzahl an Daten verarbeitet werden (Krebs & Hagenweiler, 2022, S. 13). Bei Deep-Learning-Modellen werden grosse Mengen an Inputdaten (vgl. Abbildung 3) berechnet. Diese Kalkulationen haben dabei keine logische oder physikalische Basis wie beispielsweise in der Mathematik. Zwar wird ein Output geliefert, dieser basiert jedoch auf keiner Semantik. Dadurch geht die Erklärbarkeit der Algorithmen verloren, wodurch das Blackbox-Problem entsteht (Christen et al., 2020, S. 86–87).

3.2 Blackbox-Problem

Während die ersten KI-Systeme transparent und dadurch interpretierbar waren, tritt in neueren Systemen zunehmend das Blackbox-Problem auf (Barredo Arrieta et al., 2020, S. 2). Dieses Problem beschreibt die Gefahr, dass Entscheidungsprozesse von Algorithmen für menschliche Entscheidungsträger nicht nachvollziehbar sind (Dwivedi et al., 2023, S. 10; Heim & Gerth, 2023, S. 266). Auch wenn Blackbox-Modelle aktuell noch bessere Ergebnisse liefern als Whitebox-Modelle (erklärbare Modelle) ist die Transparenz nicht gewährleistet (Heim & Gerth, 2023, S. 184). Das Fehlen dieser Transparenz ist eine der grössten Herausforderungen im KI-Bereich, da dadurch nicht sicher ist, ob Entscheide der KI aus ethischer Sicht korrekt sind (Langer & Weyerer, 2020, S. 230; Pohlink & Fischer, 2021, S. 157). Eine Möglichkeit, um diese Transparenz herzustellen, wird im Kapitel 5.1 aufgezeigt.

Das Blackbox-Problem tritt vor allem bei Anwendungen des Deep Learnings auf, da diese Anwendungen über zahlreiche Inputdaten und Verbindungen durch mehrere Schichten verfügen (Jörgens et al., 2020, S. 150) und sich ab einem bestimmten Punkt selbst programmieren (Knight, 2017, S. 5). Dabei ist nicht mehr möglich, zu verstehen, wie sich die Modelle verhalten und wie die Ergebnisse der Outputs erreicht werden (Christen et al., 2020, S. 90). Der Charakter solcher Modelle ist dabei algorithmisch und nicht analytisch wie beim Menschen. Auch wenn das Ergebnis präzise oder korrekt sein kann, ist der Weg dahin ein anderer als beim menschlichen Denken. Es liegt lediglich eine Input-Output-Beziehung (Abbildung 7) vor. Dazwischen können tausende Neuronen verbunden in hunderten Schichten liegen (Heim & Gerth, 2023, S. 266; Knight, 2017, S. 7).

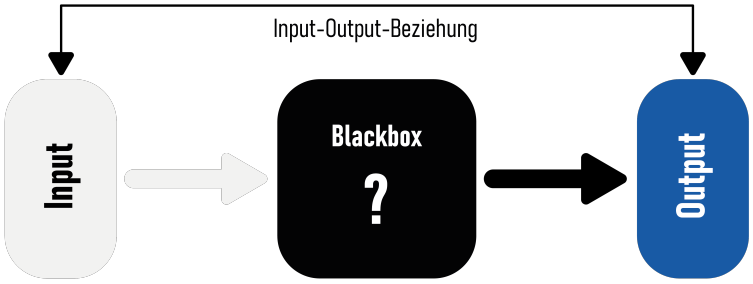

Abbildung 7: Input-Output-Beziehung im Blackbox-Modell, (Eigene Darstellung)

Eine grosse Problematik beim Blackbox-Modell liegt in der Datenverarbeitung, denn die Daten werden von einer Maschine nicht gleich verstanden und eingeordnet wie von einem Menschen. Aus diesem Grund kann es sein, dass ein Merkmal für das Modell als deutlich relevanter eingestuft wird als von einem Menschen (Orwat, 2019, S. 82). Zudem werden Merkmale automatisch miteinander in Verbindung gesetzt. Wenn deshalb beispielsweise in einem Datensatz des Recruitings das Merkmal «Geschlecht» ausgeblendet wird, kann durch andere Merkmale trotzdem darauf geschlossen werden. Eine Verhinderung eines Bias durch das Weglassen von Merkmalen ist daher zu kurz gedacht. Wie ein Bias entsteht und welche Arten davon bestehen, wird im folgenden Kapitel beschrieben.

3.3 Bias

Bei einem Bias handelt es sich um eine algorithmische Verzerrung, die dazu führt, dass Outputs und Entscheidungen eines KI-Algorithmus diskriminierend0F[1] sein können (Jörgens et al., 2020, S. 137). Ein häufiger Grund fehlerhafter Ausgaben einer KI ist eine qualitativ schlechte Datengrundlage, denn die Lernfähigkeit einer KI hängt stets von der Qualität und Menge der Daten ab (Langer & Weyerer, 2020, S. 225). Bias können jedoch auch durch Fehler im Machine-Learning-Modell entstehen (Ferrara, 2023, S. 2). Laut Jörgens et al. (2020, S. 142–152) lassen sich die Ursachen eines Bias im Machine Learning in zwei Kategorien unterteilen: «Dateninduzierter Bias» und «Modellinduzierter Bias». Dabei lassen sich diese zwei Kategorien nach Mehrabi et al. (2022, S. 4–8) noch in diverse Unterklassen einteilen.

3.3.1 Dateninduzierter Bias

Zahlreiche Probleme bei Machine-Learning-Algorithmen ergeben sich aus den Eigenschaften von Trainingsdaten, die nicht repräsentativ oder unvollständig sind (Ferrara, 2023, S. 3). Vorurteile, die während des Trainings auftreten, werden erlernt und in der KI-Anwendung reproduziert. Der Begriff einer «rein objektiv handelnden Maschine», die im bekannten Werk «Trust in numbers» von Porter (1995) erläutert wird, ist deshalb im Falle von KI-Anwendungen nicht mehr geeignet (Jörgens et al., 2020, S. 142). Mehrabi et al. (2022, S. 4–7) erläutern unterschiedliche Arten von dateninduziertem Bias. Diese werden wenn möglich direkt anhand eines Beispiels aufgezeigt.

- **Measurement-Bias**: Eine solche Verzerrung kann auftreten, wenn die falschen Merkmale ausgewählt werden oder eine falsche Gewichtung vorgenommen wird.
- **Omitted-Variable-Bias**: Dieser Bias tritt auf, wenn relevante Variablen aus dem Modell ausgelassen werden oder gar nicht zur Verfügung stehen.
- **Representation-Bias**: Ein Repräsentations-Bias tritt auf, wenn bestimmte Teile von Daten (beispielsweise Teile einer Population) in einem Gesamtdatensatz untervertreten sind. Dieser Bias entstand beim Beispiel von Amazon in Kapitel 3.6 und ist typisch für KI-Anwendungen im Recruiting (Pessach & Shmueli, 2021, S. 1). Durch

[1] Disclaimer zu den Begriffen «Diskriminierung»›, «Bias» und «Fairness»: Wenn in dieser Thesis von Diskriminierung gegenüber Frauen gesprochen wird, dann ist damit eine Verzerrung durch einen Bias gemeint, der Frauen benachteiligt. Es bestehen zahlreiche weitere Diskriminierungsdimensionen wie die Herkunft, das Alter oder die Hautfarbe sowie verschiedene Arten der Diskriminierung, worauf im Kontext dieser Arbeit nicht weiter eingegangen wird. Wenn von Fairness gesprochen wird, dann ist damit das Fehlen einer Diskriminierung oder eines Bias gemeint. Die Begriffe «Bias», «Diskriminierung» und «Unfairness» werden in der Literatur zu Algorithmen oft synonym verwendet, auch wenn diese inhaltlich nicht exakt gleichbedeutend sind (Pessach & Shmueli, 2021, S. 2).

diesen Bias entstehen verzerrte Verallgemeinerungen gegenüber unterrepräsentierten Gruppen (Langer & Weyerer, 2020, S. 227).

- **Aggregation-Bias**: Wenn aus Beobachtungen einer Gesamtpopulation Schlüsse auf ein Individuum gezogen werden, kann ein Aggregations-Bias entstehen.
- **Sampling-Bias**: Dieser Bias ist dem Repräsentations-Bias ähnlich und kann entstehen, wenn eine Stichprobe nicht zufällig durchgeführt wird und daher aus dieser nicht auf eine andere Population geschlossen werden kann.
- **Longitudinal-Data-Fallacy**: Bei Datensätzen, die über einen längeren Zeitraum analysiert werden, sollte nicht ein Querschnitt aller Daten verglichen werden, sondern es sollten die gleichen Daten über die Zeit verglichen werden. Sonst kann ein Irrtum bei Längsschnittdaten entstehen.
- **Linking-Bias:** Dieser Bias kann entstehen, wenn Netzwerkattribute (beispielsweise eine Verbindung oder Interaktion in einem Datennetzwerk) eines Users voneinander abweichen und aus diesem Grund dessen Verhalten falsch darstellen.
- **Association-Bias**: Eine Verzerrung durch Assoziation kann auf auftreten, wenn bestimmte Merkmale in eine ursächliche Wirkungsbeziehung gestellt werden, obwohl der Zusammenhang nicht kausal ist. Beispielsweise lassen höhere Gehälter bei Männern nicht automatisch auf die Leistungsfähigkeit schliessen, sondern weisen komplexere Gründe auf (Langer & Weyerer, 2020, S. 227).

3.3.2 Modellinduzierter Bias

Nicht nur die Eingabedaten, sondern auch die technische Umsetzung der Machine-Learning-Algorithmen kann einen Bias erzeugen, beispielsweise bei KI-Tools im Recruiting (Rebstadt et al., 2022, S. 495), da die Daten für den Algorithmus lediglich eine Abfolge von Zahlen sind und keine zusammenhängenden und kulturell erlernten Informationen. Dabei passt sich das Modell beim Lernen dem Problem der Daten an. Der Algorithmus sucht dafür die beste Verbindung der Eingabedaten mit den gewünschten Ausgabedaten. Wie die Daten miteinander verbunden werden, kann für Menschen unlogisch sein. Dieses Blackbox-Problem wurde bereits mehrmals erläutert und entsteht modellinduziert (Jörgens et al., 2020, S. 146–151). Mehrabi et al. (2022, S. 7) erläutern folgende Arten des modellinduzieren Bias:

- **Algorithmic-Bias**: Von einem algorithmischen Bias wird gesprochen, wenn der Bias in den Eingabedaten nicht vorhanden war und nur durch das Algorithmus-Design entstanden ist.

- **User-Interaction-Bias:** Dieser Bias kann beispielsweise im Web entstehen, einerseits durch das Interagieren des Users auf einer Webseite (Langer & Weyerer, 2020, S. 228) und andererseits durch das Interface selbst. Dieser Bias wird in zwei Kategorien unterteilt:
    - **Presentation-Bias:** Wenn bestimmte Funktionalitäten, wie beispielsweise Buttons, auf einer Website immer anders dargestellt werden, ist das Userverhalten auf dieser Website möglicherweise nicht mehr repräsentativ und es kann ein Bias entstehen.
    - **Ranking-Bias**: Bei einer Suche im Web werden höher gerankte Ergebnisse eher geklickt, wodurch diese laufend als zunehmend relevanter angesehen werden. Dieser Bias entstand beispielsweise beim Fall von TaskRabbit und Fiverr.
- **Popularity-Bias**: Daten oder Informationen, die bekannter sind, werden eher gezeigt, wodurch sie erneut an Bekanntheit gewinnen. Dies sollte jedoch kein Hinweis auf eine gute Qualität der Daten sein. Ein Popularity-Bias entsteht beispielsweise durch Fake-Reviews oder Social Bots. Diese Art von Bias kann auch als **Confirmation-Bias** verstanden werden, bei dem bestehende Informationen trotz Bias wegen ihrer Popularität eher als richtig angesehen werden (Langer & Weyerer, 2020, S. 228)
- **Emergent-Bias**: Ein emergenter Bias entsteht erst, nachdem User einige Zeit mit einem Algorithmus interagiert haben, wodurch sich ein Bias herausbildet, beispielsweise durch kulturelle Werte oder Entwicklungen in einer Gesellschaft.
- **Evaluation-Bias**: Durch das Verwenden ungleichmässiger Benchmark-Daten kann ein Evaluation-Bias entstehen. Wie die Algorithmen dabei mit den Datensätzen lernen, wird in Kapitel 4.1.2 erläutert. Dieser Bias trat (wie der Repräsentations-Bias) beim Fall von Amazon auf, da die Benchmark-Daten überwiegend männlich und deshalb unausgeglichen waren.

Da Ergebnisse einer KI je nach Anwendungsfall wieder als Inputdaten dienen, kann ein Kreislaufeffekt von sich selbst verstärkendem Bias entstehen. Dieser wird in Abbildung 8 aufgezeigt (Langer & Weyerer, 2020, S. 223).

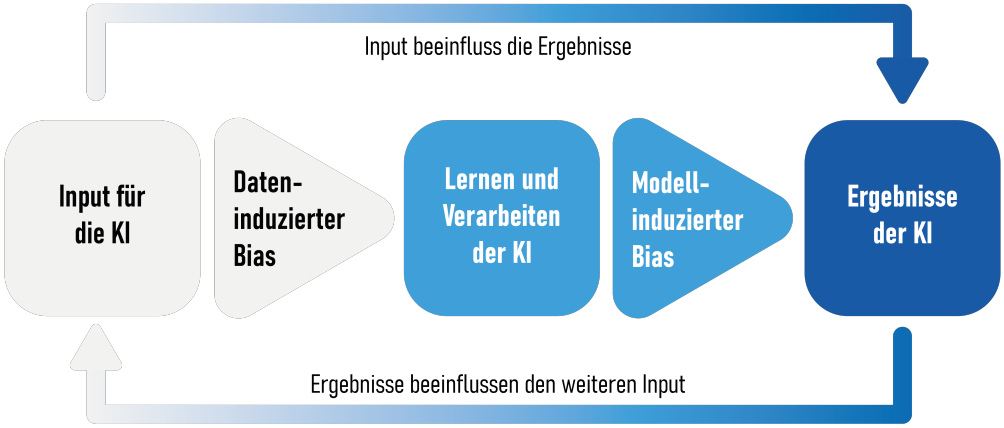

Abbildung 8: Ursachen für einen Bias in Algorithmen, (Eigene Darstellung angelehnt an Langer und Weyerer, 2020, S. 224)

3.3.3 Weitere Bias

Mehrabi et al. (2022, S. 8) definieren zudem eine weitere Art von Bias, die ausserhalb des in Abbildung 8 gezeigten Kreislaufs stattfindet, aber einen Einfluss auf die Inputdaten nimmt. Ein Beispiel hierfür ist ein **Historical Bias**, bei dem die Daten aufgrund sozialer oder anderer Ungleichheiten bereits verzerrt sind, oder ein **Population-Bias**, der durch eine Ungleichverteilung bestimmter Personengruppen in bestimmten Kontexten entsteht. Dieser Bias kann zu einem **Representation-Bias** (siehe Kapitel 4.3.1) in den Eingabedaten einer KI führen.

Wie zu erkennen ist, können in KI-Anwendungen durch die Daten sowie durch das System selbst diskriminierende Bias entstehen. Diese können dabei verschiedene Formen annehmen, deren Ursprung durch die Blackbox-Problematik schwer zu interpretieren ist.

Im Kapitel 3 wurden der Einsatz von KI im Recruiting sowie konkrete Beispiele von aufgetretenem Bias aufgezeigt. In Kapitel 4 wurde dargelegt, wie eine KI funktioniert und wie dabei ein solcher Bias entstehen kann. Da bei zahlreichen Machine-Learning-Verfahren die Gefahr einer Blackbox besteht, sind die Identifikation von Bias und deren Behebung teilweise nicht möglich, weshalb eine Whitebox ein passender Ansatz ist. Zudem kann der Ursprung eines Bias unterschiedlich sein, weshalb über den gesamten Entwicklungsprozess eines Systems sowie darüber hinaus Massnahmen dagegen vorgenommen werden sollten. Im nächsten Kapitel werden solche Massnahmen beschrieben.

4 Massnahmen gegen einen Geschlechterbias

Durch den verstärkten Einsatz von KI im Recruiting (siehe Kapitel 3) steigen die Anforderungen an die Robustheit und Fairness dieser Technologie, um das Vertrauen in solche Systeme zu stärken (Pohlink & Fischer, 2021, S. 156). Ein zentraler Schritt zur Stärkung der Fairness ist die Korrektur von Bias. Die meisten Massnahmen setzen jedoch eine nachvollziehbare KI voraus, denn ohne zu verstehen, wie eine KI funktioniert, ist es schwierig, sie diskriminierungsfrei zu gestalten (Barredo Arrieta et al., 2020, S. 46). Um dieses Blackbox-Problem zu beheben, eignet sich der Whitebox-Ansatz, der auch «Explainable AI» (XAI) genannt wird (Loyola-González, 2019, S. 154101). Da XAI die Voraussetzung zahlreicher weiterer Massnahmen ist, wird diese Massnahme in diesem Kapitel zu Beginn beschrieben.

4.1 Erklärbare KI (Whitebox-Ansatz)

XAI ist keine spezifische Massnahme in der Entwicklung eines KI-Systems, sondern ein Oberbegriff, der alle Methoden und Forschungen umfasst, um KI erklärbar zu machen (Burkart & Huber, 2021, S. 247). Da Entscheidungen zunehmend von KI übernommen werden und damit die Abhängigkeit von der Technologie steigt, wird die Transparenz und Interpretierbarkeit solcher Systeme zunehmend bedeutender (Burkart & Huber, 2021, S. 245; Dwivedi et al., 2023, S. 1). Aufgrund der Bedeutung solcher KI-Systeme ist XAI als Forschungsgebiet in den letzten Jahren relevanter geworden (Pohlink & Fischer, 2021, S. 157; Thalmann et al., 2022, S. 3) und massgeblich für die zukünftige Entwicklung der KI im Allgemeinen verantwortlich (Barredo Arrieta et al., 2020, S. 1).

Das Ziel einer XAI ist es, den Nutzerinnen und Nutzern einer KI aufzeigen zu können, auf welcher Grundlage die KI welche Entscheidungen getroffen hat und welche Gewichtung ein Merkmal (beispielsweise «Geschlecht») dabei hatte. Dadurch können einerseits das Vertrauen in ein KI-System und andererseits die Qualität des Systems selbst erhöht werden (Burkart & Huber, 2021, S. 300).

Für eine XAI sprechen verschiedene Gründe. Die erklärbare KI hilft zum einen den Entwickelnden und zum anderen den Anwenderinnen und Anwendern von KI-Systemen. Den Entwickelnden hilft sie dabei, die Funktionsweise des KI-Systems zu verstehen (Dwivedi et al., 2023, S. 2). So können Verzerrungen und damit eine Diskriminierung beim Training des Systems erkannt und behoben werden. Darüber hinaus kann die Kausalität der unterschiedlichen Gewichtungen der Merkmale verstanden werden, wodurch verhindert werden kann, dass die falschen Merkmale zur Entscheidung beitragen

(Barredo Arrieta et al., 2020, S. 2). Dies ermöglicht eine ethisch faire Entscheidungsfindung (Burkart & Huber, 2021, S. 249). Den Anwendenden von KI-Systemen hilft XAI dabei, den Output eines Systems verstehen und nachvollziehen zu können (Burkart & Huber, 2021, S. 261). Im Falle des Recruitings wären die Anwendenden einerseits die Bewerberinnen und Bewerber, andererseits die Personalverantwortlichen.

4.1.1 Ablauf der XAI-Entwicklung

Der Ablauf der XAI wird nach Dwivedi et al. (2023, S. 3) in fünf Phasen unterteilt: Lernen, Testen, Verstehen, Einsetzen und Erklären. Die Phase «Lernen» erfolgt dabei während der Entwicklung des Systems. Die Phasen «Testen» und «Verstehen» erfolgen in der Testphase und die Phasen «Einsetzen» und «Erklären» in der Produktion – also, wenn das KI-System veröffentlicht wurde und neue Outputs produziert. Die Phase «Lernen» kann dabei mit den Lernmethoden des Machine-Learnings aus dem Kapitel 4.1.2 verglichen werden. Auch die Phasen «Testen» und «Einsetzen» kommen beim üblichen Machine- und Deep Learning zum Einsatz. Wie in Abbildung 9 erkennbar ist, sind bei der XAI zusätzlich die Phasen «Verstehen» und «Erklären» vorhanden. In beiden Phasen nehmen unterschiedliche Stakeholder einen Einfluss auf die erklärbare KI. Während in der Phase des Verstehens die Entwickelnden, Theoretikerinnen und Theoretiker (Fachpersonen aus der Domäne) sowie Data-Scientists mitwirken, sind an der Phase des Verstehens die User, Konsumentinnen und Konsumenten oder Anbietende des Systems beteiligt. Zudem ist die XAI in dieser Phase für Regulierungsbehörden relevant, insbesondere in Anwendungen, die rechtliche Folgen haben können. Auf solche Behörden als Stakeholder wird jedoch in dieser Arbeit nicht weiter eingegangen (Dwivedi et al., 2023, S. 3)

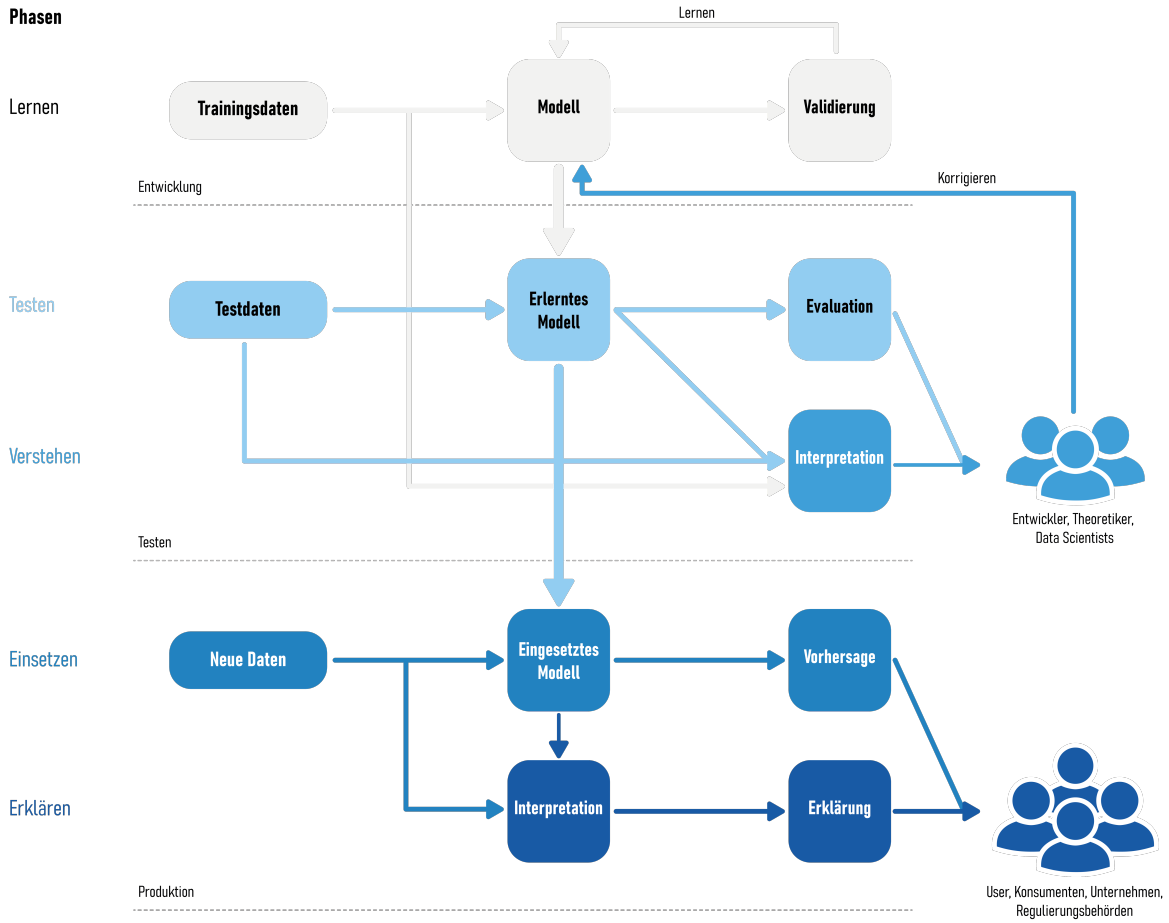

Abbildung 9: Ablauf Erklärbare KI, (Eigene Darstellung angelehnt an Dwivedi, 2023, S. 3-5)

Ein zentraler Aspekt der Phase «Verstehen» ist das Training des Modells mit Testdaten und die damit verbundene Qualitätssicherung. Dabei werden die Ergebnisse von den Stakeholdern interpretiert und es wird überprüft, ob das System wie gewünscht funktioniert. Die Interpretation erfolgt auf der Ebene der Merkmale, deren Wechselwirkungen untereinander analysiert werden. In dieser Phase werden mögliche Verzerrungen entdeckt, die korrigiert werden können. Bei der «Erklärung» nehmen andere Stakeholder einen Einfluss auf das System und überprüfen, ob die Anwendung auch mit realen Daten noch die gewünschten Ergebnisse liefert. Auf dieser Ebene müssen die Daten für alle User interpretierbar und begründbar sein. Für Unternehmen ist die Erklärbarkeit gegenüber den Usern in diesem Schritt von grosser Bedeutung, da durch die Begründbarkeit der Outputs Diskriminierungsvorwürfen entgegengewirkt werden kann (Dwivedi et al., 2023, S. 3–4).

4.1.2 Blackbox vs. Whitebox

Die erklärbare KI kann in zwei Kategorien eingeteilt werden. Die Einteilung hängt dabei vom Zeitpunkt ab, wann das System erklärbar gemacht wurde. Es wird zwischen der intrinsischen (Ante-hoc) sowie der nachträglichen (Post-hoc) Erklärbarkeit unterschieden (Abbildung 10). Bei der Ante-hoc-Methode wird die Erklärbarkeit direkt während der Entwicklung in die Struktur des Systems eingebaut. Daher wird von einer Whitebox gesprochen (Loyola-González, 2019, S. 154096). Beim Post-hoc-Ansatz wird die Erklärbarkeit erst nach der Entwicklung eines Blackbox-Systems vollzogen (Du et al., 2019, S. 69).

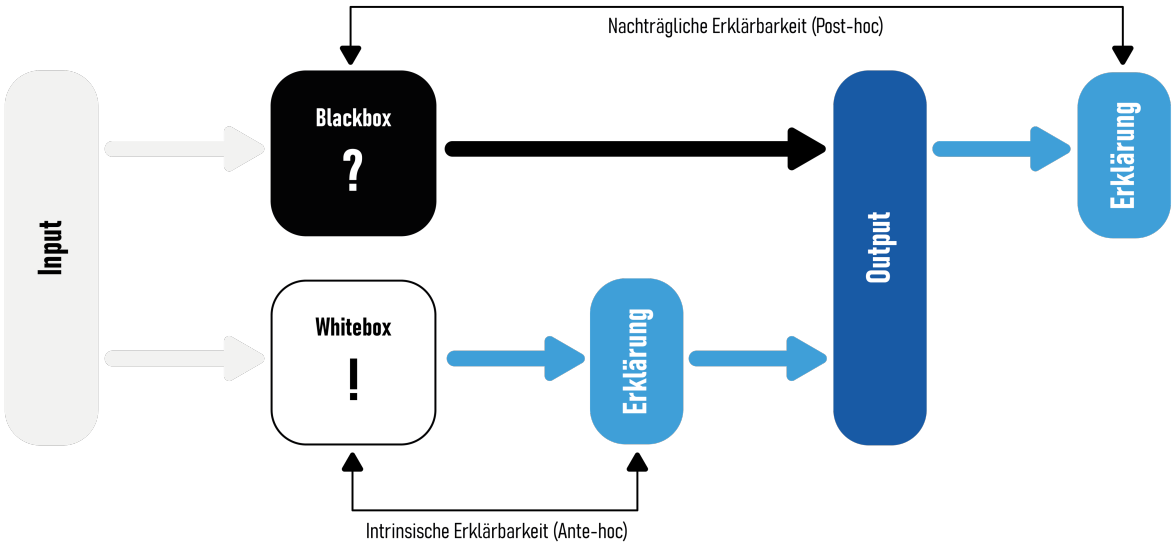

Abbildung 10: Blackbox vs. Whitebox, (Eigene Darstellung)

Beide Ansätze sind in der KI-Entwicklung verbreitet und haben unterschiedliche Stärken. Ante-hoc-Systeme erzielen ihre Ergebnisse auf Basis von Modellen, die für den Menschen leichter interpretierbar sind. Die Erklärbarkeit wird bereits zu Beginn der Entwicklung berücksichtigt und in das System eingebaut (Burkart & Huber, 2021, S. 256). Fehler wie eine Verzerrung können deshalb früh in der Entwicklung erkannt und behoben werden (Barredo Arrieta et al., 2020, S. 38). Hierfür eignen sich verschiedene Modelle. Bekannte Beispiele sind Decision-Trees oder Rule-based Models (Burkart & Huber, 2021, S. 257–258; Loyola-González, 2019, S. 154102). Durch die Ähnlichkeit zur menschlichen Sprache und Logik ist die Erklärbarkeit für Expertinnen und Experten gegeben (Loyola-González, 2019, S. 154096). Der grosse Nachteil von Whitebox-Modellen liegt in der Performance des Systems und der Genauigkeit der Ergebnisse (Du et al., 2019, S. 70). Dieser Trade-off zwischen Genauigkeit und Erklärbarkeit wird in Abbildung 11 dargestellt.

Blackbox-Modelle liefern im Gegensatz zu Whitebox-Modellen präzisere Outputs (Du et al., 2019, S. 70; Heim & Gerth, 2023, S. 184; Loyola-González, 2019, S. 154096). Die

Erklärbarkeit ist dagegen aufwändiger, da ein zusätzliches Modell programmiert werden muss, das die Erklärbarkeit möglich macht (Du et al., 2019, S. 69). Da die Interpretierbarkeit erst nach der Kreierung des Modells erreicht wird, gehen die Vorteile des Ante-hoc-Ansatzes verloren (vgl. Kapitel 5.1.1), beispielsweise das Beheben eines Bias während des Verstehens des Modells (Burkart & Huber, 2021, S. 256). Eine nachträgliche Korrektur ist zudem deutlich aufwändiger und erfordert ein tiefes Verständnis des Aufbaus und der internen Zusammenhänge des Algorithmus (Jörgens et al., 2020, S. 150). Dwivedi et al. (2023, S. 2) weisen darauf hin, das sich eine bessere Performance eines KI-Modells nicht rechtfertigen lässt, wenn sich dadurch ein Bias einschleichen kann, der nicht erklärt werden kann.

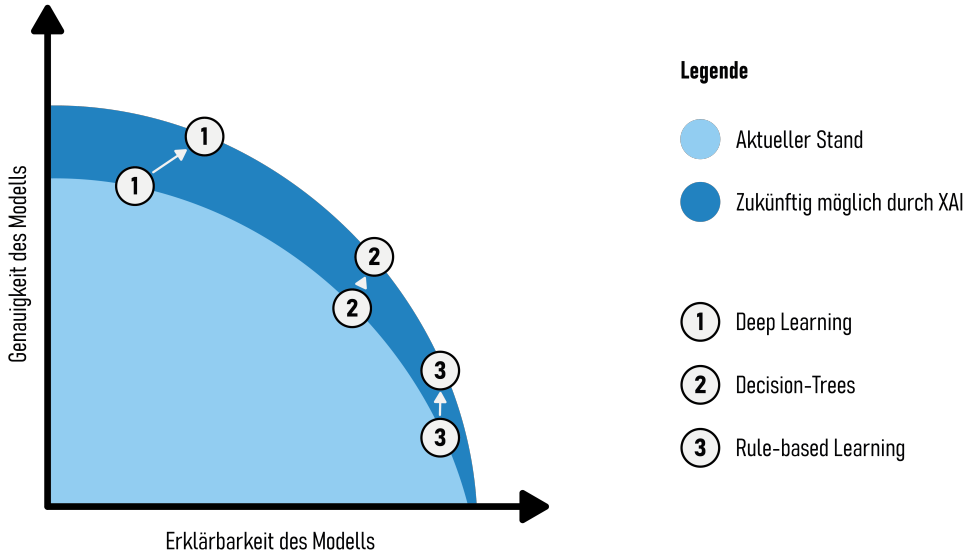

Abbildung 11: Trade-off zwischen Genauigkeit und Erklärbarkeit bei KI-Modellen, (Eigene Darstellung angelehnt an Bar-redo Arrieta et al., 2020, S. 31; Gunning und Aha, 2019, S. 46)

Bestimmte Anforderungen an eine KI lassen sich nicht gleichzeitig erreichen. So besteht ein Zielkonflikt zwischen Genauigkeit und Fairness (Christen et al., 2020, S. 108). Wird beispielsweise die Fairness für bestimmte Gruppen berücksichtigt, kann an anderer Stelle des Modells die Genauigkeit verlorengehen. Daher ist bei der Entwicklung eines Systems abzuwägen, auf welche Verzerrungen ein Fokus gelegt werden soll und inwieweit dabei die Genauigkeit zu berücksichtigen ist (Ferrara, 2023, S. 9). Darüber hinaus kann sich das Problem ergeben, die Diskriminierung quantifizierbar und damit korrigierbar zu machen (Jörgens et al., 2020, S. 149). In Abbildung 11 ist der Unterschied zwischen einem Deep-Learning-Model (präzise, aber kaum erklärbar) und erklärbaren Modellen wie Decision-Trees oder dem Rule-based Learning ersichtlich.

Die erklärbare KI bietet die Grundlage, auf der weitere Werte wie die Fairness oder Ethik einer KI aufbauen können (Barredo Arrieta et al., 2020, S. 46). Die Erklärbarkeit eines

Modells reicht jedoch nicht aus, damit dieses fair und diskriminierungsfrei handelt. Die XAI bietet die Möglichkeit, ein Modell zu verstehen. Erst wenn ein Verständnis für ein KI-Modell vorhanden ist, können Massnahmen zur Verbesserung getroffen werden (Fry, 2019, S. 31; Pohlink & Fischer, 2021, S. 157). Aus diesem Grund werden nachfolgend weitere Massnahmen erläutert. Massnahmen hierzu werden in der Literatur häufig in die Kategorien «Pre-Processing», «In-Processing» und «Post-Processing» eingeteilt (Barredo Arrieta et al., 2020, S. 39; Mehrabi et al., 2022, S. 13–14; Samek et al., 2019). Dabei ist der betrachtete Prozess jener vom Input bis zum Output des KI-Systems (vgl. dazu Abbildung 7 und Abbildung 8). Die diesem Prozess vorgelagerten Massnahmen werden als Pre-Processing und die nachgelagerten Massnahmen als Post-Processing bezeichnet.

4.2 Pre-processing Massnahmen

Bei Pre-Processing-Massnahmen werden die Daten so aufbereitet, dass im anschliessenden Lernen möglichst keine Verzerrung auftreten kann. Der Bias soll also direkt im ersten Prozessschritt eliminiert werden (Barredo Arrieta et al., 2020, S. 39; Z. Chen, 2023, S. 145), denn eine mangelhafte Datenqualität bietet ein grosses Fehlerpotenzial und dadurch Konsequenzen für die Stakeholder aus Abbildung 9 (Heim & Gerth, 2023, S. 23, 121; Pohlink & Fischer, 2021, S. 159). Neben einer guten Qualität ist auch eine grosse Menge an Daten notwendig, damit das Modell richtig lernen kann (de Laat, 2018, S. 530; Heim & Gerth, 2023, S. 122; Krebs & Hagenweiler, 2022, S. 11). Zudem müssen die Daten aktuell sein und den Kontext, in dem sie später verwendet werden, korrekt abbilden (Jörgens et al., 2020, S. 145). Eine Möglichkeit besteht darin, die Trainingsdaten so anzupassen, dass sie die diskriminierten Gruppen angemessen repräsentieren. Dies kann z. B. durch **Oversampling**, **Undersampling** oder **Synthetic-Data-Generation** geschehen. Sowohl beim Oversampling als auch beim Undersampling wird eine Verzerrung eingeführt, indem mehr Stichproben aus einer Klasse als aus einer anderen ausgewählt werden, um ein bestehendes Ungleichgewicht in den Daten auszugleichen. Bei der Synthetic-Data-Generation werden aus bestehenden Daten künstliche Daten erstellt, um eine gewünschte Gewichtung bestimmter Datenmerkmale zu erreichen (Ferrara, 2023, S. 7). Eine andere Möglichkeit besteht durch **Adversarial Debiasing**. Dabei kann eine KI robuster gegenüber einer Verzerrung durch Inputdaten programmiert werden. Es wird ein Modell trainiert, welches beispielsweise eine Vorhersage zum Attribut «Geschlecht» vornimmt. Danach wird die Gewichtung der Merkmale je nach Output angepasst (Zhang et al., 2018, S. 1). Herausforderungen bei Pre-Processing-Massnahmen sind, dass diese oft aufwendig sind, da sie eine intensive Arbeit an den Daten und Merkmalen benötigen.

Zudem ist spezifisches Fachwissen für den Anwendungsbereich und die möglichen Verzerrungen nötig, da die Massnahmen sonst nicht effektiv sein können (Ferrara, 2023, S. 8). Wenn die Massnahmen an den Trainingsdaten erfolgreich durchgeführt wurden, kann ein dateninduzierter Bias (vgl. Kapitel 4.3.1) verhindert werden.

4.3 In-processing Massnahmen

In-Processing-Massnahmen erfolgen während des Trainings des Modells und sollen während des Lernprozesses die Diskriminierung reduzieren. Dabei werden während des Lernprozesses Änderungen am Algorithmus vorgenommen. Die Vermeidung eines Bias kann in dieser Phase durch die Wahl von Modellen, die transparent sind und Fairness priorisieren, oder durch die Kombination verschiedener Modelle erreicht werden (Ferrara, 2023, S. 8). Mögliche Modelle hierzu wurden in Kapitel 5.1 erwähnt. Hinzu kommt die Möglichkeit der Bestrafung des Algorithmus bei einer diskriminierenden Ausgabe (ähnlich wie beim Reinforcement-Learning aus Kapitel 4.1.2.3). Massnahmen während des Lernprozesses sind bekannt dafür, einen guten Trade-off zwischen Genauigkeit und Erklärbarkeit zu erreichen (vgl. Kapitel 5.1.2). Zudem kann eine verbesserte Performance erreicht werden, da die Genauigkeit sowie die Erklärbarkeit direkt in den Algorithmus eingebaut werden. Der Nachteil dabei liegt jedoch darin, dass diese Mechanismen eng an das KI-Modell gebunden sind (Pessach & Shmueli, 2021, S. 3).

4.4 Post-processing Massnahmen

Post-Processing-Massnahmen finden nach dem Lernprozess des Modells statt und können deshalb mit dem Post-hoc-Ansatz (Abbildung 10) verglichen werden. Die Massnahmen greifen weder in die Inputdaten noch in das Modell beim Lernprozess ein und eignen sich für Blackbox-Modelle, da bei diesen Pre-Process- und In-Process-Massnahmen nicht möglich sind (Barredo Arrieta et al., 2020, S. 39). Die Outputs aus dem Modell werden kritisch betrachtet und auf eine Verzerrung überprüft. Dadurch kann ein möglicher Bias im Nachhinein korrigiert werden (Ferrara, 2023, S. 7). Bei Post-Process-Massnahmen können Hürden eintreten. Einerseits ist die nachträgliche Kontrolle komplex und bedarf einer grossen Menge neuer Daten, um die bisherigen Outputs zu überprüfen. Andererseits ist es nicht zwingend möglich, den Ursprung einer Verzerrung nachzuweisen – insbesondere, wenn es sich um eine Blackbox handelt. In der Post-Processing-Phase sind die Daten bereits aufbereitet und durch den Algorithmus gelaufen. Im Nachhinein zu rekonstruieren, wie und wo ein Bias entstanden ist, kann aufwendig oder

unmöglich sein. Wenn Verzerrungen entdeckt werden, sind sie zudem schwieriger zu messen und zu quantifizieren (Ferrara, 2023, S. 8–9).

Die bisher vorgestellten Massnahmen bezogen sich in erster Linie auf den Entwicklungsprozess einer KI-Anwendung. Daneben können weitere Massnahmen eingesetzt werden, die laufend ergriffen werden können und nicht direkt mit der Entwicklung der Algorithmen zusammenhängen, sondern eher für die Anwendenden (Recruiter) relevant sind. Diese werden im folgenden Unterkapitel aufgezeigt.

4.5 Laufende Massnahmen

In den folgenden Unterkapiteln werden Massnahmen erläutert, die Anwendenden eines fertig entwickelten Systems vornehmen können, um eine Verzerrung zu vermeiden. Diese sind relevant ab dem Moment, ab dem die Personalverantwortlichen ein KI-Tool aus der eigenen Entwicklungsabteilung oder aus externer Entwicklung einsetzen wollen. Dabei sind die Massnahmen spezifisch für Recruiter ausgearbeitet.

4.5.1 Zusammenarbeit zwischen Menschen und KI definieren

Die KI sollte im Recruiting als Unterstützung und nicht als Ersatz gesehen werden. Dabei kann die Technologie Recruiter entlasten und Zeit für Aufgaben freisetzen, die besser von Menschen erledigt werden können (Employ & JOBVITE, 2023, S. 18). Welche Aufgaben die KI bereits gut ersetzen kann, wird in der nächsten Massnahme erläutert. Die KI ist dem menschlichen Recruiter in quantitativen Arbeiten, die gut messbar und auswertbar sind, durch eine hohe Rechenleistung überlegen (Heim & Gerth, 2023, S. 265). Fry (2019) meint, diese Schwächen sollten anerkannt werden: «*Wir müssen […] unsere eigenen Schwächen anerkennen, wir müssen unser Bauchgefühl hinterfragen und uns die Gefühle gegenüber den Algorithmen, die uns umgeben, bewusst machen.*» (S. 37).

Wenn es jedoch um zwischenmenschliche Werte geht, ist es nicht möglich, ausschliesslich der Ergebnisse einer KI zu betrachten. Beispielsweise kann die Ausstrahlung einer Person in einem Bewerbungsgespräch nicht vollständig digital erfasst werden. Das gesamte sinnliche Erleben eines Menschen kann nicht digital abgebildet werden (Spiekermann, 2021, S. 88–93). Da zwischenmenschliche Werte für Bewerbende von hoher Bedeutung sind (IU Internationale Hochschule, 2022, S. 7), sollte während des Bewerbungsprozesses zu einem bestimmten Zeitpunkt stets Kontakt zu einem Menschen bestehen. Zudem sollten endgültige Entscheidungen stets von einer Person getroffen werden. Ein Algorithmus kann dabei unterstützend wirken und bei der Entscheidungsfindung

helfen, indem er grosse Datenmengen sammelt und filtert (siehe Abbildung 12). Am Ende sollte jedoch ein Mensch mit Vetorecht die Entscheidung überprüfen (Fry, 2019, S. 33).

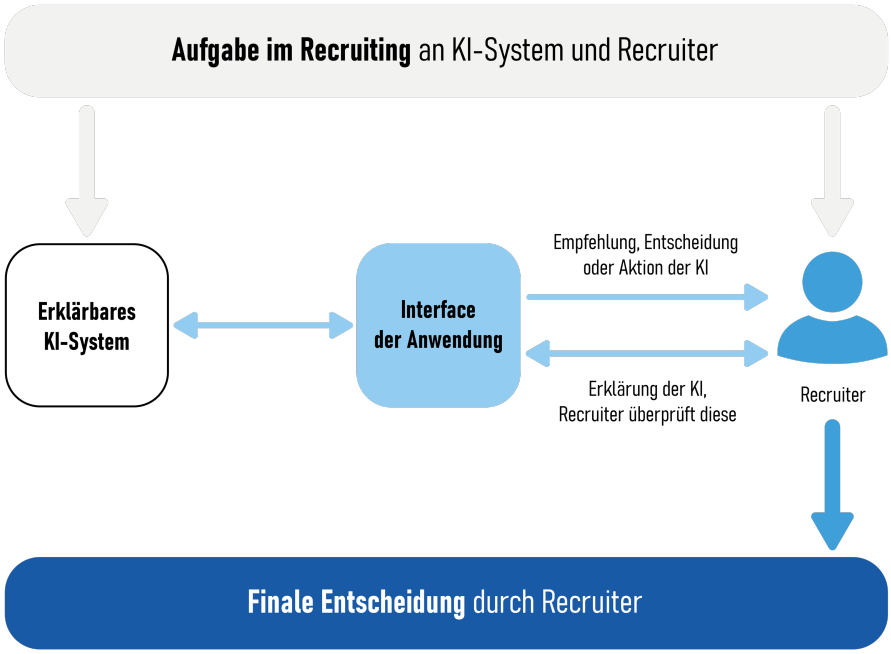

Abbildung 12: Entscheidungsframework durch XAI, (Eigene Darstellung angelehnt an Gunning und Aha, 2019, S. 50)

Ein weiterer Aspekt in der Zusammenarbeit und Anwendung von KI sind diverse Teams. Aktuell sind Entwickler von KI-Anwendungen primär männlich, was dazu führen kann, dass Maschinen so programmiert werden, dass sie eine maskuline Denkweise reproduzieren (Hassanien et al., 2021, S. 311). Einerseits in der Entwicklung, andererseits in der Anwendung von Tools kann durch ein diverses Team dazu beigetragen werden, eine KI fairer zu gestalten und zu verwenden (Ferrara, 2023, S. 7).

4.5.2 Festlegen spezifischer Anwendungsfälle

Nachdem erläutert wurde, wie die Rekrutierenden mit der KI zusammenarbeiten sollen, werden nachfolgend Aufgaben erwähnt, die sich für diese Zusammenarbeit besonders eignen. Für einige Arbeiten im Recruiting eignen sich KI-Tools gegenwärtig besser als für andere. Insbesondere weniger komplexe Tasks können von einer KI präziser und transparenter durchgeführt werden. Daher lohnt es sich, den genauen Anwendungsbereich zu definieren und die Risiken darin abzuschätzen, bevor ein Tool eingesetzt wird (Onlyfy, 2023, S. 32). Im Kapitel 3.3 wurden Aufgaben aufgezeigt, die aktuell bereits von KI übernommen werden. Dabei bieten die folgenden Aufgaben eine geringe Komplexität: Erstellen und Optimieren von Stellenanzeigen, Job-Profile-Matching sowie die Active-Sourcing Kommunikation (Onlyfy, 2023, S. 13). Thalmann et al. (2022, S. 1) nennen als

besonders geeignete Anwendungen die Zielgruppenansprache, das Screening sowie die Vorauswahl der Bewerbenden. Folglich kann gesagt werden, dass sich KI-Tools vor allem in den früheren Stufen des Bewerbungsprozesses (Abbildung 1) eignen, da die Komplexität der Aufgaben dort noch eher gering ist.

4.5.3 Ethische Prinzipien und Richtlinien

Um im Recruiting einen Rahmen zu haben, an dem eine Orientierung bei ethischen Fragen erfolgen kann, lohnt es sich, genaue Richtlinien zu erstellen. Diese können dabei helfen, einen sicheren und fairen Umgang mit der KI zu gewährleisten (Ferrara, 2023, S. 7; Onlyfy, 2023, S. 10). Dabei wären beispielsweise für das Recruiting branchenübergreifende Guidelines möglich, damit zahlreiche Unternehmen davon profitieren können (Wilke & Bendel, 2022, S. 663). Solche Guidelines sollten bei der Entwicklung eines eigenen KI-Tools beachtet werden. Beim Kauf eines Systems von einem Drittanbieter sollte dieses Unternehmen zur Einhaltung verpflichtet werden, denn insbesondere bei der externen Beschaffung von KI-Lösungen ist die Gefahr einer Blackbox gross (Pohlink & Fischer, 2021, S. 159, 162).

Um einen Anhaltspunkt zu haben, können HR-Abteilungen oder Branchenverbände bereits gegenwärtig auf bestehende Richtlinien und ethische Prinzipien zurückgreifen. Barton und Pöppelbuss (2022, S. 476–479) stellen in ihrem Modell sechs Prinzipien vor, die zu einer ethischen KI führen sollen. Die Prinzipien sind: Wohltätigkeit, Transparenz, Nicht-Boshaftigkeit, Autonomie, Gerechtigkeit und Datenschutz. Mit «Wohltätigkeit» ist gemeint, dass eine KI dem Menschen dienen soll und zum Guten verwendet wird. Die «Transparenz» wurde im Kapitel 5.1 bereits erklärt. «Nicht-Boshaftigkeit» soll Schäden an Nutzerinnen und Nutzern durch das System vermeiden. Durch die «Autonomie» soll erreicht werden, dass der Mensch als Nutzer in der Entscheidungsfindung nicht abhängig vom System wird. Mit «Gerechtigkeit» ist die Fairness gemeint, die in Kapitel 4.3 erläutert wird. Durch den «Datenschutz» sollen abschliessend die Privatsphäre sowie das Dateneigentum der User geschützt werden.

Fjeld et al. (2020, S. 5) definieren acht Prinzipien: Privacy, Accountability, Safety and Security, Transparency and Explainability, Fairness and Non-Discrimination, Human Control of Technology, Professional Responsibility sowie Promotion of Human Values. Wie in Abbildung 13 zu erkennen ist, sind beide Modelle grösstenteils deckungsgleich, auch wenn die Prinzipien teilweise anders benannt werden. Bei den Prinzipien von Fjeld et al. (2020, S. 5) kommt jedoch der Punkt der «Professional Responsibility» dazu. Mit diesem Prinzip wird die Bedeutung aller Beteiligten bei der Entwicklung und

Implementierung von KI-Systemen erwähnt und es wird betont, dass ihre Professionalität und Integrität erforderlich sind, um langfristige Auswirkungen zu berücksichtigen.

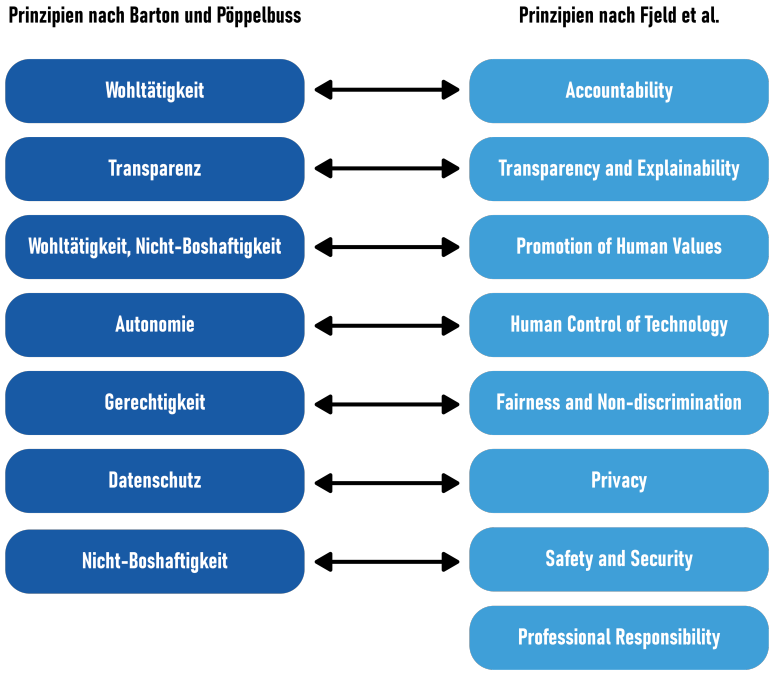

Abbildung 13: Vergleich ethischer Prinzipien, (Eigene Darstellung angelehnt an Barton und Pöppelbuss, 2022. S. 476-479; Fjeld et al., 2020, S.5)

Diese beiden Richtlinien dienen als Beispiele, da sie die meisterwähnten Werte solcher Prinzipien aus der Literatur abdecken und aufzeigen, dass die Inhalte darin oft ähnlich sind. Es finden sich zahlreiche weitere solcher Richtlinien. Jobin et al. (2019, S. 7) analysierten 84 solcher ethischen Guidelines zu KI und verglichen diese miteinander. Dabei zeigte sich, dass sich die Inhalte häufig überschneiden. Beispielsweise wurde die Transparenz als Prinzip in 73 der 84 Guidelines erwähnt. Aus solchen Prinzipien können anschliessend Richtlinien erstellt werden. Ein Beispiel solcher Richtlinien bieten die «Ethik-Leitlinien Für Eine Vertrauenswürdige KI» der Europäischen Kommission (Generaldirektion Kommunikationsnetze, Inhalte und Technologien (Europäische Kommission), 2019).

4.5.4 Schulungen und Awareness

Um das HR-Personal auf den Umgang mit KI-Tools vorzubereiten, empfiehlt es sich, Schulungen durchzuführen. Diese können von internen Personen mit entsprechender Expertise durchgeführt werden oder es werden externe Expertinnen und Experten hinzugezogen (Onlyfy, 2023, S. 10; Wilke & Bendel, 2022, S. 664). Zudem ist während der Einführung eines solchen Tools sowie während des Betriebs jederzeit Awareness zu den Risiken eines Bias zu schaffen (Wilke & Bendel, 2022, S. 663).

4.5.5 Transparenz gegenüber Bewerbenden schaffen

Wie ein transparentes System aufgebaut werden kann, wurde im Kapitel 5.1 anhand von XAI aufgezeigt. Nebst der Transparenz des Systems selbst sollte auch der Bewerbungsprozess gegenüber den Bewerbenden transparent gestaltet werden. Es wird empfohlen, den Bewerbenden proaktiv zu kommunizieren, in welchen Schritten des Bewerbungsprozesses KI eingesetzt wird (Onlyfy, 2023, S. 10). Bewerbende sollten wissen, wie das KI-System funktioniert und was mit den eingegebenen Daten geschieht. Zudem ist eine rechtzeitige und verständliche Information bedeutend, da sie das Vertrauen stärken kann (Thalmann et al., 2022, S. 1–3; Wilke & Bendel, 2022, S. 664). Da eine Blackbox-KI diese Erklärung nicht ermöglicht, bringt sie im Recruitment sowohl bei den Recruitern als auch bei den Bewerbenden ein grosses Akzeptanzproblem mit sich (Thalmann et al., 2022, S. 3).

4.5.6 Kontinuierliche Überwachung

Abschliessend ist eine kontinuierliche Überwachung der vorher erläuterten Massnahmen zwingend. Zur Überwachung können auch Feedbackschleifen und Audits eingesetzt werden (Ferrara, 2023, S. 7). Dabei sollten das KI-System und dessen Outputs überprüft und gegebenenfalls angepasst werden (Pohlink & Fischer, 2021, S. 159). So kann die Anwendung kontinuierlich optimiert werden, wobei sichergestellt werden kann, dass die hohen ethischen Anforderungen an die Rekrutierung eingehalten werden können (Basler de Roca, 2023, S. 5–6; Onlyfy, 2023, S. 11).

5 Diskussion

Das Ziel dieser Bachelorthesis ist die Erarbeitung konkreter Massnahmen, um KI-Tools in der Personalrekrutierung frei von geschlechtsspezifischer Diskriminierung zu gestalten. In Form einer Literaturarbeit wurde dabei der aktuelle Einsatz von KI im Recruiting sowie die Technologie an sich näher betrachtet. Da bei der Analyse der Literatur festgestellt wurde, dass primär die Stakeholder der KI-Entwicklung und der Personalrekrutierung Einfluss auf die Ergebnisse solcher KI-Tools haben, wurden die Massnahmen spezifisch für diese zwei Anspruchsgruppen definiert. Um die Massnahmen für die Entwicklung herzuleiten, wurde eine Analyse der Technologie KI gemacht und für die Massnahmen der Personalrekrutierung eine Betrachtung des aktuellen Einsatzes von KI in ihrer Tätigkeit. Die Massnahmen werden zudem in eine zeitliche Abfolge eingeteilt, je nachdem ob sie vor, während oder nach der Entwicklung eines Systems getätigt werden sollten.

5.1 Interpretation der Ergebnisse

In den folgenden Kapiteln sollen die Forschungsfragen beantwortet werden. Dafür werden zuerst die beiden Unterfragen beantwortet, da diese die Basis bilden, um die Hauptforschungsfrage beantworten zu können.

5.1.1 Beantwortung Unterfrage 1

Aufgrund des Kapitels 3 kann die Unterfrage 1 (*Wie wird KI aktuell bei Unternehmen im Recruiting-Prozess eingesetzt?*) folgendermassen beantwortet werden: Aktuell wird KI nur in einem kleinen Teil der Personalabteilungen eingesetzt. Dabei wird die Technologie von der Ausschreibung eines Stelleninserats bis hin zur Bewertung eines Bewerbungsgesprächs über den gesamten Rekrutierungsprozess hinweg verwendet (Black & van Esch, 2020, S. 218). Die Gründe für den Einsatz von KI im Recruiting sind primär eine Effizienzsteigerung und eine damit verbundene Kosteneinsparung (Wilke & Bendel, 2022, S. 655). Recruiter können sich durch diese Zeiteinsparung auf andere relevante Arbeiten fokussieren (Guenole & Feinzig, 2018, S. 7). Trotz der Vorteile sind die Implementierung und der Betrieb solcher KI-Systeme mit Herausforderungen verbunden, insbesondere hinsichtlich der Vermeidung einer diskriminierenden Verzerrung und der Gewährleistung der Transparenz der Entscheidungen. Aus diesem Grund warnen Fachpersonen vom Einsatz von KI im Recruiting. Bewerbende sind grösstenteils eher negativ gegenüber solchen Tools eingestellt und wünschen sich den Kontakt mit einer realen

Person (IU Internationale Hochschule, 2022). Die Recruiter sind sich zwar den Vorteilen durch KI in ihrem Arbeitsalltag bewusst, sehen aber auch die Gefahr eines möglichen Jobverlusts durch die Technologie (Personio, 2023b, S. 23).

5.1.2 Beantwortung Unterfrage 2

Als Basis der Unterfrage 2 (*Wie funktionieren KI-Tools, sodass dabei ein Geschlechterbias entsteht?*) dient das Kapitel 4, in dem die Technologie der KI genauer analysiert wurde. Die Unterfrage kann wie folgt beantwortet werden: Die Technologie der KI basiert auf Algorithmen und Neural Networks. Algorithmen sind dabei Handlungsanweisungen für die Datenverarbeitungen des Computers, während Neural Networks künstliche Neuronen sind, die dem menschlichen Gehirn nachempfunden sind (Krebs & Hagenweiler, 2022, S. 8–14). Eine KI kann auf verschiedene Arten lernen, die als Machine-Learning zusammengefasst werden. Dabei bestehen drei Arten des Machine-Learnings: Suvervised Learning, Unsupervised Learning und Reinforcement-Learning (Heim & Gerth, 2023, S. 120). Eine zentrale Problematik ist das sogenannte Blackbox-Problem, bei dem die Entscheidungsprozesse der Algorithmen für menschliche User nicht nachvollziehbar sind. Diese fehlende Transparenz führt dazu, dass die Fairness nicht überprüft und verbessert werden kann (Barredo Arrieta et al., 2020, S. 2). Zudem kann eine Diskriminierung unentdeckt bleiben oder verstärkt werden (Dwivedi et al., 2023, S. 10). Die zahlreichen Bias können in zwei Hauptkategorien unterteilt werden (Jörgens et al., 2020, S. 142–152). Ein dateninduzierter Bias kann entstehen, wenn Trainingsdaten nicht repräsentativ oder unvollständig sind. Ein Beispiel ist der Representation-Bias, der beim Case von Amazon auftrat und entstehen kann, wenn bestimmte Daten in einem Gesamtdatensatz untervertreten sind. Nebst dem dateninduzierten kann der modellinduzierte Bias auftreten. Dieser entsteht nicht durch die Eingabedaten, sondern erst durch die technische Umsetzung einer KI. Die beiden Arten von Bias können sich gegenseitig beeinflussen und verstärken (Langer & Weyerer, 2020, S. 223).

5.1.3 Beantwortung Forschungsfrage

In Kapitel 5 wurden Massnahmen gegen einen Bias in KI-Systemen aufgezeigt. Die Forschungsfrage (***Welche Massnahmen können ergriffen werden, um einen Geschlechterbias durch den Einsatz von KI-Tools im Recruiting zu vermeiden?***) soll in diesem Kapitel beantwortet werden. Dabei werden die Massnahmen zusammengefasst als Handlungsempfehlungen erläutert und danach eingeteilt, ob sie für die Entwicklung oder für das Recruiting relevant sind.

5.1.3.1 Handlungsempfehlungen Entwicklung

Da Entwicklerinnen und Entwickler von KI-Rekrutierungstools in den Phasen vom Pre-Processing bis zum Post-Processing einen Einfluss auf eine faire KI nehmen, sind für sie die Massnahmen aus den Kapiteln 5.1 bis 5.4 relevant.

Eine der wesentlichen Massnahmen ist das Erreichen einer erklärbaren KI (Whitebox). Zahlreiche weitere Massnahmen für eine faire KI setzen diese Erklärbarkeit voraus (Barredo Arrieta et al., 2020, S. 46) und erst dadurch wird eine ethische Entscheidungsfindung im Recruiting möglich (Burkart & Huber, 2021, S. 249). Da Whitebox-Modelle grundsätzlich jedoch weniger genaue Ergebnisse liefern, ist ein passender Trade-off zwischen Genauigkeit und Erklärbarkeit abzuwägen (Ferrara, 2023, S. 9). Im **Pre-Processing** sollen die Daten so aufbereitet werden, dass beim anschliessenden Lernen des Modells keine Verzerrung auftreten kann (Barredo Arrieta et al., 2020, S. 39). Um eine repräsentative Datenmenge mit genug hohem Anteil an weiblichen Daten zu erreichen, können Techniken wie Oversampling, Undersampling und Synthetic-Data-Generation verwendet werden. **In-Processing**-Massnahmen umfassen die Möglichkeiten, die während des Trainings des Modells ergriffen werden können. Dabei können Lernmodelle ausgewählt werden, die transparent sind und Fairness priorisieren. In dieser Phase kann ein guter Trade-off zwischen Genauigkeit und Transparenz erreicht werden (Ferrara, 2023, S. 7–8). **Post-Process**-Massnahmen finden nach dem Lernprozess statt und eignen sich vor allem für Blackbox-Modelle gut (Barredo Arrieta et al., 2020, S. 39). Dabei werden die Outputs des KI-Systems auf einen Bias überprüft. Der Aufwand in dieser Phase ist hoch und das Nachweisen des Ursprungs eines Bias ist nicht immer möglich (Ferrara, 2023, S. 7–9). Aus diesem Grund empfehlen sich eher Pre- oder In-Processing-Massnahmen, die ein System nachhaltiger diskriminierungsfrei gestalten und nicht nur die Ergebnisse daraus korrigieren.

5.1.3.2 Handlungsempfehlungen Recruiting

Normalerweise wenden Recruiter ein KI-Tool erst an, nachdem es fertig entwickelt wurde. Aus diesem Grund sind für sie die Massnahmen aus Kapitel 5.4 und 5.5 relevant, die ein entwickeltes System voraussetzen. Einige der Massnahmen für die Recruiter setzen kein spezifisches KI-Tool voraus, sondern sollten allgemein für den Umgang mit KI definiert werden.

Die Zusammenarbeit mit der KI sollte genau definiert werden. Die KI soll dabei das Recruiting in denjenigen Aufgaben unterstützen, in denen die Leistung gut auswertbar ist und sie dem Menschen aufgrund ihrer hohen Rechenleistung überlegen ist. Aufgaben, die eine zwischenmenschliche Komponente besitzen, sowie relevante Entscheide sollten

stets von einem Recruiter getätigt werden. Zudem sollten spezifische Anwendungsfälle für den Einsatz der Technologie definiert werden. Die HR-Abteilung sollte genau festlegen, in welchen Phasen des Recruitings die KI welche Aufgaben übernehmen kann. Dabei empfiehlt es sich, die Technologie eher in den früheren Stufen des Bewerbungsprozesses einzusetzen (Thalmann et al., 2022, S. 1). Einerseits sind die Aufgaben dort weniger komplex und andererseits ist die Akzeptanz der Bewerbenden gegenüber der KI höher (IU Internationale Hochschule, 2022).

Um einen fairen Umgang mit KI-Anwendungen im Recruiting gewährleisten zu können, sollten Richtlinien definiert werden (Ferrara, 2023, S. 7). Dabei können die HR-Abteilung oder die Branchenverbände auf bestehende Prinzipien aus der Literatur zur KI-Ethik zurückgreifen. Um die Massnahmen im Team umsetzen zu können, sollten Recruiter für einen fairen Umgang mit KI-Tools geschult und über mögliche Gefahren informiert werden (Wilke & Bendel, 2022, S. 664). Des Weiteren sollte gegenüber den Bewerbenden Transparenz über den Einsatz von KI bestehen. Um Vertrauen zu schaffen, sollen Bewerbende über den Einsatz von KI rechtzeitig und verständlich informiert werden (Thalmann et al., 2022, S. 1–3; Wilke & Bendel, 2022, S. 664). Alle genannten Massnahmen sollten nicht nur einmal umgesetzt und anschliessend nicht mehr beachtet werden. Um den hohen ethischen Ansprüchen des Recruitings gerecht zu werden, sollen alle Massnahmen kontinuierlich überwacht werden. Bei Bedarf müssen Massnahmen zur Verbesserung ergriffen werden (Ferrara, 2023, S. 7; Pohlink & Fischer, 2021, S. 159).

In Abbildung 14 werden alle Massnahmen nochmals zusammengefasst aufgezeigt. Dabei ist ersichtlich, dass die Massnahmen für die Entwicklung (XAI sowie Modellentwicklung) auch zeitgleich verlaufen können. Die Ante-hoc-Massnahmen finden dabei Pre- und In-Process statt, während die Post-hoc-Massnahmen Post-Process erfolgen. Die laufenden Massnahmen für das Recruitment werden erst nach der Entwicklung durchgeführt.

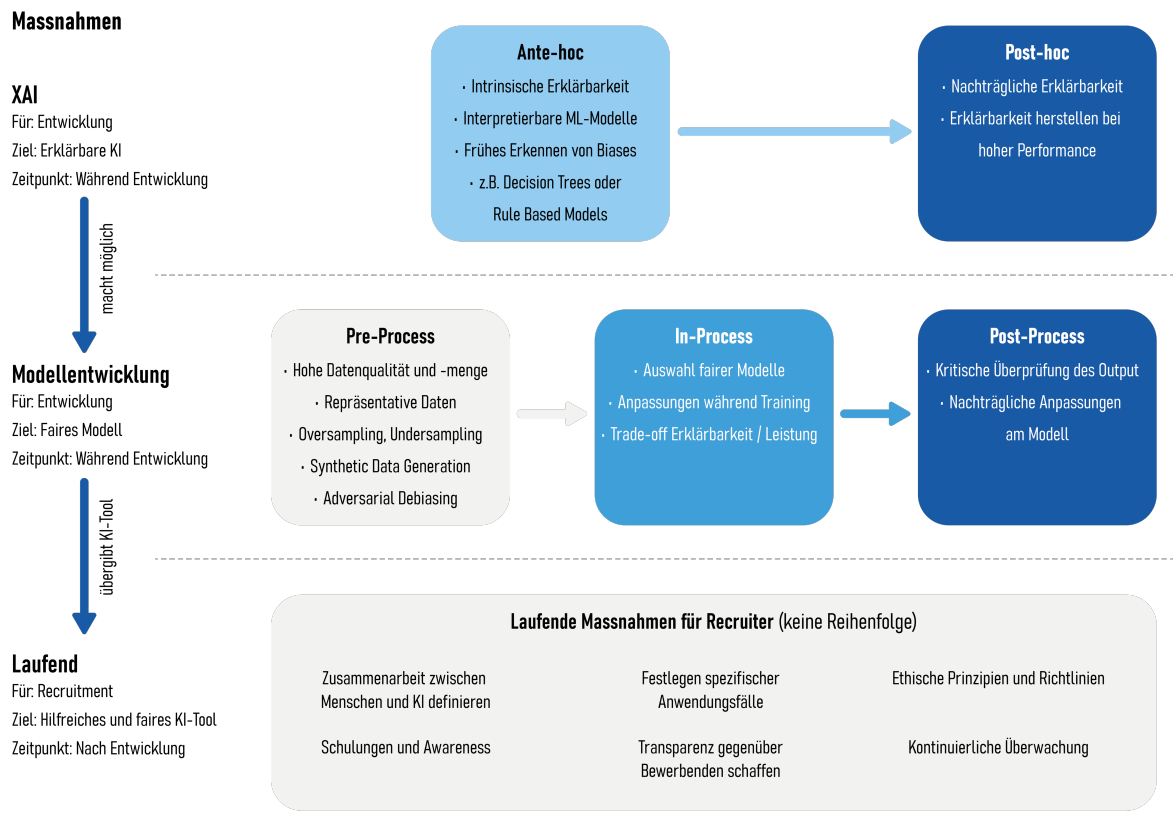

Abbildung 14: Zusammengefasste Massnahmen für die Entwicklung und die Personalrekrutierung, (Eigene Darstellung)

5.2 Limitationen und Stärken

In dieser reinen Literaturarbeit lag der Fokus auf Fachliteratur zu den Themen Recruiting und KI. Durch eine andere Methodik, beispielsweise durch qualitative Interviews mit KI-Fachpersonen oder Recruitern, hätten wertvolle Insights direkt aus der Praxis mit hoher Aktualität gewonnen werden können. Der aktuelle Standpunkt aus der Branche hätte betrachtet werden können. Zudem hätte bereits analysiert werden können, welche Massnahmen aktuell gegenüber Verzerrungen in den Algorithmen vorgenommen werden. Diese Arbeit war daher durch die reine Literatur und Analyse bestehender Fälle limitiert. Die in der Literatur erwähnten Fälle von aufgetretenem Genderbias sind ausschliesslich von internationalen Grossunternehmen. Für ein differenzierteres Bild wäre dementsprechend ein Fall eines kleineren Unternehmens hilfreich gewesen. Da die Handlungsempfehlungen für zwei Zielgruppen (Entwicklung und Rekrutierung) erarbeitet wurden, litt die Tiefe der Empfehlungen. Bei einem Fokus auf einen Stakeholder hätten die Handlungsempfehlungen umfangreicher, konkreter und dadurch wertvoller sein können. Insbesondere bei den Empfehlungen für die Entwicklung hätten durch einen grösseren Umfang mehr technische Möglichkeiten aufgezeigt werden können. Die genannten Massnahmen, beispielswiese das Adversarial Debiasing, konnten so nicht umfangreich genug

beschrieben werden. Ausserdem wurden in dieser Arbeit keine rechtlichen Aspekte bezüglich des geltenden Rechts bei der Personalrekrutierung oder des Datenschutzes bei KI-Systemen thematisiert. Durch die Berücksichtigung dieser Thematik könnten weitere Massnahmen aufgezeigt werden. So könnten die definierten Massnahmen, beispielsweise das Erstellen ethischer Richtlinien, auf eine rechtliche Basis gestützt werden.

Die Bachelorthesis schafft einen Beitrag zum Bewusstsein für geschlechterspezifische Bias-Probleme durch KI. Da die Phasen des typischen Rekrutierungsprozesses aufgezeigt wurden, konnten Handlungsempfehlungen entlang des gesamten Prozesses aufgezeigt werden. Die Einteilung der Massnahmen nach Pre-Process, In-Process und Post-Process ermöglicht es, dass diese den Stakeholdern später strukturierter zugeordnet werden konnten. Die Aufteilung zeigt zudem die Komplexität und die Wechselwirkungen zwischen den Prozessschritten auf. Durch die Darstellung der Entwicklung des Recruitings wurde deutlich, warum gegenwärtig KI eingesetzt wird und welche Gründe dafür sprechen. Dabei wurden nicht nur die Gefahren und negativen Seiten der Technologie beleuchtet, sondern auch die positiven Aspekte. In einer Zeit, in der KI-Tools einen Boom erleben und zahlreiche Personen solche Anwendungen ohne ethische Bedenken nutzen, konnte eine allgemeine Sensibilisierung für das Thema erreicht werden. Durch die Beschreibung der Ursachen und Arten von Bias und die verschiedenen Fälle aus der Praxis konnte aufgezeigt werden, dass zahlreiche Einflussfaktoren auf einen Bias in einem Algorithmus einwirken können. Daneben wurde dargelegt, wie anfällig solche KI-Systeme sein können. Da die Inhalte der Arbeit teilweise in reiner Textform nicht leicht verständlich sind, konnte durch selbst erstellte, einheitliche Grafiken eine Hilfestellung geschaffen werden, die den Text visuell unterstützt oder verständlicher macht. Insbesondere bei technischeren Themen wie XAI oder der Entwicklung der Whitebox bieten die Grafiken einen grossen Mehrwert zum Geschriebenen.

6 Fazit

Mit dieser Bachelorthesis konnten Massnahmen für einen diskriminierungsfreien Einsatz von KI im Recruiting definiert werden. Die Ergebnisse daraus stellen den aktuellen Stand aus der Fachliteratur dar. Es wurden verschiedene Massnahmen definiert, die einerseits von Entwicklerinnen und Entwicklern und andererseits von Recruitern umgesetzt werden können. Die hohe Anzahl und die Vielfalt der Massnahmen zeigen bereits, wie komplex das Thema ist, und dass an verschiedenen Stellen zu unterschiedlichen Zeitpunkten Handlungsbedarf besteht. Aufgrund der grossen Fortschritte, die KI aktuell macht, unterstreicht diese Bachelorthesis die Bedeutung zukünftiger Forschung in diesem Bereich. Es wird zentral sein, die zunehmend komplexer werdenden Algorithmen zu verstehen und dadurch Massnahmen für einen fairen Umgang damit definieren zu können. Die Ergebnisse dieser Arbeit können jedoch bereits dabei helfen, einen faireren Umgang gegenüber Frauen durch KI-Tools im Recruiting zu gewährleisten.

Jedoch ist festzuhalten, dass durch Algorithmen keine sozialen und gesellschaftlichen Ungleichstellungen gelöst werden. Auch wenn im In-Process und Post-Process durch neue Möglichkeiten ein Bias kontinuierlich verringert werden kann, werden die Daten (also die Pre-Process-Inputs) nach wie vor mit der Aussenwelt und den gesellschaftlichen Strukturen verknüpft sein. Die Ergebnisse einer KI zeigen Probleme auf, die in der Gesellschaft und Kultur verankert sind. Eine Anpassung der Algorithmen für gerechtere und ausgewogenere Ergebnisse ist daher eher eine Symptombekämpfung als eine endgültige Lösung für tiefer liegende Probleme. Wie im Kapitel 4.3 erklärt wurde, entsteht ein Bias im Machine-Learning stets in Wechselwirkung mit der realen Welt und den Menschen, die darin agieren. Es ist erfreulich, wenn die in Kapitel 5 beschriebenen Massnahmen einen Geschlechterbias im Recruiting verhindern oder zumindest minimieren könnten. Jedoch ist eine Verzerrung, die Frauen benachteiligt, nur eine von zahlreichen solcher Verzerrungen, während das Recruiting nur ein Anwendungsfall einer hohen Anzahl von Fällen ist. In den kommenden Jahren wird durch die grosse Popularität der KI eine Vielzahl neuer Tools und Möglichkeiten mit neuen Problemen und Verzerrungen entstehen. Auch hier werden nach einiger Zeit Optimierungsmöglichkeiten bestehen. Um allerdings eine faire Nutzung solcher Werkzeuge für alle zu gewährleisten, bedarf es einer gesellschaftlichen und politischen Entwicklung.

7 Literaturverzeichnis

Adelmann, L., & Wiedmer, J. (2017). *Der Einsatz von Künstlicher Intelligenz in der Rekrutierung*. https://wwz.unibas.ch/fileadmin/user_upload/wwz/00_Professuren/Beckmann_Personal_und_Organisation/Lehre/Digital_Transformation/Der_Einsatz_von_Kuenstlicher_Intelligenz_in_der_Rekrutierung_Adelmann_und_Wiedmer.pdf

Ballestrem, J. G., Bär, U., Gausling, T., Hack, S., & Von Oelffen, S. (2020). *Künstliche Intelligenz: Rechtsgrundlagen und Strategien in der Praxis*. Springer Fachmedien Wiesbaden. https://doi.org/10.1007/978-3-658-30506-2

Bardy, G.-R., Gyöngyösi, N., & Mölleney, M. (2022). Künstliche Intelligenz in der Personalauswahl: Debatte. *Personal Schweiz*, *2022*(März), 26–27. https://doi.org/10.21256/zhaw-25407

Barredo Arrieta, A., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., Garcia, S., Gil-Lopez, S., Molina, D., Benjamins, R., Chatila, R., & Herrera, F. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, *58*, 82–115. https://doi.org/10.1016/j.inffus.2019.12.012

Barton, M.-C., & Pöppelbuß, J. (2022). Prinzipien für die ethische Nutzung künstlicher Intelligenz. *HMD Praxis der Wirtschaftsinformatik*, *59*(2), 468–481. https://doi.org/10.1365/s40702-022-00850-3

Basler de Roca, R. (2023). *Recruiting Now*. *02*, 4–6.

Black, J. S., & van Esch, P. (2020). AI-enabled recruiting: What is it and how should a manager use it? *Business Horizons*, *63*(2), 215–226. https://doi.org/10.1016/j.bushor.2019.12.001

Böhm, S., Linnyk, O., Jäger, W., & Teetz, I. (2021). KI im Recruiting: Anwendungsfelder, Entwicklungsstand und Anwendungsbeispiele aus der Praxis. In T. Barton & C. Müller (Hrsg.), *Künstliche Intelligenz in der Anwendung: Rechtliche Aspekte, Anwendungspotenziale und Einsatzszenarien* (S. 195–218). Springer Fachmedien. https://doi.org/10.1007/978-3-658-30936-7_11

Burkart, N., & Huber, M. F. (2021). A Survey on the Explainability of Supervised Machine Learning. *Journal of Artificial Intelligence Research*, *70*, 245–317. https://doi.org/10.1613/jair.1.12228

Chen, L., Ma, R., Hannák, A., & Wilson, C. (2018). Investigating the Impact of Gender on Rank in Resume Search Engines. *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, 1–14. https://doi.org/10.1145/3173574.3174225

Chen, Z. (2023). Collaboration among recruiters and artificial intelligence: Removing human prejudices in employment. *Cognition, Technology & Work*, *25*(1), 135–149. https://doi.org/10.1007/s10111-022-00716-0

Christen, M., Mader, C., Čas, J., Tarik, A.-C., Bernstein, A., Braun Binder, N., Dell'Aglio, D., Fábián, L., George, D., Gohdes, A., Hilty, L., Kneer, M., Krieger-Lamina, J., Licht, H., Scherer, A., Som, C., Sutter, P., & Thouvenin, F. (2020). *Wenn Algorithmen für uns entscheiden: Chancen und Risiken der künstlichen Intelligenz* (TA-SWISS, Hrsg.; 1. Aufl.). vdf Hochschulverlag AG an der ETH Zürich. https://doi.org/10.3218/4002-9

Datta, A., Tschantz, M. C., & Datta, A. (2015). *Automated Experiments on Ad Privacy Settings: A Tale of Opacity, Choice, and Discrimination* (arXiv:1408.6491). arXiv. https://doi.org/10.48550/arXiv.1408.6491

de Laat, P. B. (2018). Algorithmic Decision-Making Based on Machine Learning from Big Data: Can Transparency Restore Accountability? *Philosophy & Technology*, *31*(4), 525–541. https://doi.org/10.1007/s13347-017-0293-z

Dike, H. U., Zhou, Y., Deveerasetty, K. K., & Wu, Q. (2018). Unsupervised Learning Based On Artificial Neural Network: A Review. *2018 IEEE International Conference on Cyborg and Bionic Systems (CBS)*, 322–327. https://doi.org/10.1109/CBS.2018.8612259

Döring, N., & Bortz, J. (2016). Forschungsstand und theoretischer Hintergrund. In N. Döring & J. Bortz, *Forschungsmethoden und Evaluation in den Sozial- und Humanwissenschaften* (S. 157–179). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-41089-5_6

Du, M., Liu, N., & Hu, X. (2019). Techniques for interpretable machine learning. *Communications of the ACM*, *63*(1), 68–77. https://doi.org/10.1145/3359786

Dwivedi, R., Dave, D., Naik, H., Singhal, S., Omer, R., Patel, P., Qian, B., Wen, Z., Shah, T., Morgan, G., & Ranjan, R. (2023). Explainable AI (XAI): Core Ideas, Techniques, and Solutions. *ACM Computing Surveys*, *55*(9), 194:1-194:33. https://doi.org/10.1145/3561048

Employ, & JOBVITE. (2023). *Automation and AI in Recruiting: Balancing the Risks and Rewards in a Modern Hiring Environment* [Whitepaper]. employ. https://web.jobvite.com/rs/328-BQS-080/images/2023-05-Jobvite-Employ-Thought-Leadership-Report-Q2-2023.pdf

Ferrara, E. (2023). Fairness and Bias in Artificial Intelligence: A Brief Survey of Sources, Impacts, and Mitigation Strategies. *Sci*, *6*(1), Article 1. https://doi.org/10.3390/sci6010003

Fjeld, J., Achten, N., Hilligoss, H., Nagy, A., & Srikumar, M. (2020). *Principled Artificial Intelligence: Mapping Consensus in Ethical and Rights-Based Approaches to Principles for AI* (SSRN Scholarly Paper 3518482). https://doi.org/10.2139/ssrn.3518482

Fry, H. (2019). *Hello world: Was Algorithmen können und wie sie unser Leben verändern* (S. Schmid, Übers.). C.H. Beck.

Generaldirektion Kommunikationsnetze, Inhalte und Technologien (Europäische Kommission). (2019). *Ethik-leitlinien für eine vertrauenswürdige KI*. Amt für Veröffentlichungen der Europäischen Union. https://data.europa.eu/doi/10.2759/22710

Guenole, N., & Feinzig, S. (2018). The Business Case for AI in HR: With Insights and Tips on Getting Started. In *IBM Publishing* [Report]. IBM Corporation. https://research.gold.ac.uk/id/eprint/33662/

Hasenbein, M. (2023). Künstliche Intelligenz und Roboter im Human Resources Bereich. In M. Hasenbein (Hrsg.), *Mensch und KI in Organisationen: Einfluss und Umsetzung Künstlicher Intelligenz in wirtschaftspsychologischen Anwendungsfeldern* (S. 85–107). Springer. https://doi.org/10.1007/978-3-662-66375-2_6

Hassanien, A. E., Haqiq, A., Tonellato, P. J., Bellatreche, L., Goundar, S., Azar, A. T., Sabir, E., & Bouzidi, D. (Hrsg.). (2021). *Proceedings of the International Conference on Artificial Intelligence and Computer Vision (AICV2021)* (Bd. 1377). Springer International Publishing. https://doi.org/10.1007/978-3-030-76346-6

Heim, L., & Gerth, S. (Hrsg.). (2023). *Entrepreneurship der Zukunft: Voraussetzung, Implementierung und Anwendung von Künstlicher Intelligenz im Rahmen datenbasierter Geschäftsmodelle*. Springer Fachmedien Wiesbaden. https://doi.org/10.1007/978-3-658-42060-4

index Research. (2023). *Barometer Personalvermittlung 2023* (Barometer Personalvermittlung). index Research. https://anzeigendaten.index.de/white-paper/barometer-personalvermittlung-2023/

IU Internationale Hochschule. (2022). *KI im Recruiting: Das denken Bewerber:innen | IU Studie*. IU – Internationale Hochschule. https://www.iu.de/forschung/studien/ki-im-recruiting-studie/

Jäger, W. (2018). „Recruiting 1.0 – 4.0": Strategien, Prozesse und Systeme im Wandel der Zeit. In C. Kochhan & A. Moutchnik (Hrsg.), *Media Management: Ein interdisziplinäres Kompendium* (S. 1–27). Springer Fachmedien. https://doi.org/10.1007/978-3-658-23297-9_1

Jares, P., & Vogt, T. (2021). Der Einsatz von KI-basierter Sprachanalyse im Bewerbungsverfahren. In I. Knappertsbusch & K. Gondlach (Hrsg.), *Arbeitswelt und KI 2030: Herausforderungen und Strategien für die Arbeit von morgen* (S. 75–82). Springer Fachmedien. https://doi.org/10.1007/978-3-658-35779-5_8

Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. *Nature Machine Intelligence*, *1*(9), 389–399. https://doi.org/10.1038/s42256-019-0088-2

Jörgens, D., Rieder, Y., & Sinzinger, F. (2020). Bias in Machine Learning und Konsequenzen für die Anwendung in der Marktforschung. In B. Keller, H.-W. Klein, A. Wachenfeld-Schell, & T. Wirth (Hrsg.), *Marktforschung für die Smart Data World* (S. 137–156). Springer Fachmedien Wiesbaden. https://doi.org/10.1007/978-3-658-28664-4_11

Kambur, E., & Yildirim, T. (2022). Changes in Human Resources Management with Artificial Intelligence. In M. Virvou, G. A. Tsihrintzis, N. G. Bourbakis, & L. C. Jain (Hrsg.), *Handbook on Artificial Intelligence-Empowered Applied Software Engineering: VOL.2: Smart Software Applications in Cyber-Physical Systems* (S. 89–102). Springer International Publishing. https://doi.org/10.1007/978-3-031-07650-3_6

Knight, W. (2017). *The Dark Secret at the Heart of AI*.

Krebs, H.-A., & Hagenweiler, P. (2022). *Datenanonymisierung im Kontext von Künstlicher Intelligenz und Big Data: Grundlagen – Elementare Techniken – Anwendung*. Springer Fachmedien. https://doi.org/10.1007/978-3-658-37588-1

Langer, P. F., & Weyerer, J. C. (2020). Diskriminierungen und Verzerrungen durch Künstliche Intelligenz. Entstehung und Wirkung im gesellschaftlichen Kontext. In M. Oswald & I. Borucki (Hrsg.), *Demokratietheorie im Zeitalter der Frühdigitalisierung* (S. 219–240). Springer Fachmedien. https://doi.org/10.1007/978-3-658-30997-8_11

Lavanchy, M. (2018). *Amazon's sexist hiring algorithm could still be better than a human*. IMD. https://www.imd.org/research-knowledge/digital/articles/amazons-sexist-hiring-algorithm-could-still-be-better-than-a-human/

Loyola-González, O. (2019). Black-Box vs. White-Box: Understanding Their Advantages and Weaknesses From a Practical Point of View. *IEEE Access*, *7*, 154096–154113. IEEE Access. https://doi.org/10.1109/ACCESS.2019.2949286

McCarthy, J., Minsky, M. L., Rochester, N., & Shannon, C. E. (1955). A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence, August 31, 1955. *AI Magazine*, *27*(4), Article 4. https://doi.org/10.1609/aimag.v27i4.1904

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2022). A Survey on Bias and Fairness in Machine Learning. *ACM Computing Surveys*, *54*(6), 1–35. https://doi.org/10.1145/3457607

Onlyfy. (2023). *KI im Recruiting: Leitfaden für den erfolgreichen Einsatz von Künstlicher Intelligenz in der Personalgewinnung* [Whitepaper]. onlyfy. https://onlyfy.com/wp-content/uploads/onlyfy-whitepaper-kuenstliche-intelligenz-im-recruiting.pdf?lid=fq3mgj40zq8p&bid=8b8cbd0e81c65e6f73da0d2d4bcaf35f91ee5118e953d64ded0d8f7752502595

Orwat, C. (2019). *Diskriminierungsrisiken durch Verwendung von Algorithmen*.

Personio. (2023a). *Keine Spielerei mehr: KI als wichtiger Business-Faktor für HR* [Whitepaper]. Personio. https://www.personio.de/hr-lexikon/kuenstliche-intelligenz-im-recruiting-ai-in-hr/

Personio. (2023b). *Schulterschluss von HR und Geschäftsführung: Wie Unternehmen zukunftssicher werden*. Personio. https://www.personio.de/ueber-uns/presse/ki-studie/

Pessach, D., & Shmueli, E. (2021). Improving fairness of artificial intelligence algorithms in Privileged-Group Selection Bias data settings. *Expert Systems with Applications*, *185*, 115667. https://doi.org/10.1016/j.eswa.2021.115667

Pohlink, C., & Fischer, S. (2021). Verantwortungsvolle und robuste KI in Unternehmen: Wie man KI-bezogene Risiken gegen Bias und Diskriminierung beherrscht. In I. Knappertsbusch & K. Gondlach (Hrsg.), *Arbeitswelt und KI 2030* (S. 155–163). Springer Fachmedien Wiesbaden. https://doi.org/10.1007/978-3-658-35779-5_16

Porter, T. M. (1995). *Trust in Numbers: The Pursuit of Objectivity in Science and Public Life*. Princeton University Press. https://doi.org/10.1515/9781400821617

Rebstadt, J., Kortum, H., Gravemeier, L. S., Eberhardt, B., & Thomas, O. (2022). Non-Discrimination-by-Design: Handlungsempfehlungen für die Entwicklung von vertrauenswürdigen KI-Services. *HMD Praxis der Wirtschaftsinformatik*, *59*(2), 495–511. https://doi.org/10.1365/s40702-022-00847-y

Reindl, C., & Krügl, S. (2023). Praktische Anwendungsgebiete von People Analytics. In C. Reindl & S. Krügl (Hrsg.), *People Analytics in der Praxis: Mit Datenanalyse zu besseren Entscheidungen im Personalmanagement* (S. 193–232). Haufe. https://doi.org/10.34157/978-3-648-15851-7_7

Samek, W., Montavon, G., Vedaldi, A., Hansen, L. K., & Müller, K.-R. (Hrsg.). (2019). *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning* (Bd. 11700). Springer International Publishing. https://doi.org/10.1007/978-3-030-28954-6

Spiekermann, S. (2021). *Digitale Ethik: Ein Wertesystem für das 21. Jahrhundert*. Droemer Taschenbuch.

Suzuki, K. (Hrsg.). (2011). *Artificial Neural Networks—Methodological Advances and Biomedical Applications*. InTech. https://doi.org/10.5772/644

Teetz, I. (2021). Künstliche Intelligenz im Recruiting. In T. Petry & W. Jäger (Hrsg.), *Digital HR: smarte und agile Systeme, Prozesse und Strukturen im Personalmanagement* (2. Auflage). Haufe Group.

Thalmann, S., Malin, C., Kupfer, C., Fleiß, J., Griesbacher, M., & Kubicek, B. (2022). Künstliche Intelligenz in der Personalauswahl: Schlussfolgerungen und Empfehlungen aus einer aktuellen Studie der Universität Graz im Auftrag des AMS Österreich. *AMS info*, *544*.

Wilke, G., & Bendel, O. (2022). KI-gestütztes Recruiting – technische Grundlagen, wirtschaftliche Chancen und Risiken sowie ethische und soziale Herausforderungen. *HMD Praxis der Wirtschaftsinformatik*, *59*(2), 647–666. https://doi.org/10.1365/s40702-022-00849-w

Wuttke, L. (2023, Mai 24). *Machine Learning vs. Deep Learning: Wo ist der Unterschied?* datasolut GmbH. https://datasolut.com/machine-learning-vs-deep-learning/

Zhang, B. H., Lemoine, B., & Mitchell, M. (2018). Mitigating Unwanted Biases with Adversarial Learning. *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, 335–340. https://doi.org/10.1145/3278721.3278779

Zuiderveen Borgesius, F. (2018). *Discrimination, artificial intelligence, and algorithmic decision-making*. https://dare.uva.nl/search?identifier=7bdabff5-c1d9-484f-81f2-e469e03e2360

Bisher erschienene Schriften

Ergebnisse von Forschungsprojekten erscheinen jeweils in Form von Arbeitsberichten in Reihen. Sonstige Publikationen erscheinen in Form von alleinstehenden Schriften.

Derzeit gibt es in den Churer Schriften zur Informationswissenschaft folgende Reihen:
Reihe Berufsmarktforschung

Weitere Publikationen

Churer Schriften zur Informationswissenschaft – Schrift 165
Herausgegeben von Wolfgang Semar
Alina Viert
Herausforderungen in der Aufbewahrung von Videospielen und ihrer Peripherie
Fragen und Antworten insbesondere zur Peripherie und zur Emulation als Lösungsansatz
Chur 2023
ISSN 1660-945X

Churer Schriften zur Informationswissenschaft – Schrift 166
Herausgegeben von Wolfgang Semar
Susanne Knöpfel
Wissenslandkarten als Grundlage für Visualisierungen im Wissensmanagement
Chur, 2023
ISSN 1660-945X

Churer Schriften zur Informationswissenschaft – Schrift 167
Herausgegeben von Wolfgang Semar
Lorena Staiger
Deep Web und Bibliotheken: Stand der Dinge
Chur, 2023
ISSN 1660-945X

Churer Schriften zur Informationswissenschaft – Schrift 168
Herausgegeben von Wolfgang Semar
Karin Mattmann
Positive Darstellungen archivarischer Tätigkeiten in Fiktion
Wie das Abbild von fiktionalem Archivpersonal in der Öffentlichkeit positiv und realistisch dargestellt werden kann
Chur, 2023
ISSN 1660-945X

Churer Schriften zur Informationswissenschaft – Schrift 169
Herausgegeben von Wolfgang Semar
Stefan Banzer
Codemigration mit ChatGPT
Evaluation von ChatGPT als Tool zur teilautomatisierten Codeübersetzung von COBOL Code zu Python Code
Chur, 2023
ISSN 1660-945X

Churer Schriften zur Informationswissenschaft – Schrift 170
Herausgegeben von Wolfgang Semar
Marion Spitz
Digitale Nudges zwischen Moral und Manipulation
Eine quantitative Inhaltsanalyse zu den Auswirkungen ethischer Aspekte auf die erforschte Wirksamkeit von digitalen Nudges
Chur, 2024
ISSN 1660-945X

Churer Schriften zur Informationswissenschaft – Schrift 171
Herausgegeben von Wolfgang Semar
Joy Walser
Erschliessungsmöglichkeiten einer Sammlung mit Records in Contexts
Entwicklung und Anwendung eines konzeptionellen Modells für die Sammlung «Pfarrer F. Tschugmell, Siegel- und Stempelsammlung»
Chur, 2024
ISSN 1660-945X

Churer Schriften zur Informationswissenschaft – Schrift 172
Herausgegeben von Wolfgang Semar
Alessio Monte
Potenzialanalyse zur Anwendung von KI-basierten
Chur, 2024
ISSN 1660-945X

Churer Schriften zur Informationswissenschaft – Schrift 173
Herausgegeben von Wolfgang Semar
Lisa Köllner
Der Familienbezug und seine Bedeutung für die Nutzung von Firmenarchiven durch Familienunternehmen am Beispiel aktuell tätiger Unternehmen
Chur, 2024
ISSN 1660-945X

Churer Schriften zur Informationswissenschaft – Schrift 174
Herausgegeben von Wolfgang Semar
Silvia Rutz
Psychologische Sicherheit in virtuellen agilen Teams
Eine explanative Analyse der Einflussfaktoren auf die psycholo-gische Sicherheit in virtuellen agilen Software-Teams
Chur, 2024
ISSN 1660-945X

Churer Schriften zur Informationswissenschaft – Schrift 175
Herausgegeben von Wolfgang Semar
Jérôme Gander
Information Governance und öffentliche Verwaltung
Definitionen, Nutzen und die Rolle der Verwaltungsarchive.
Chur, 2024
ISSN 1660-945X

Churer Schriften zur Informationswissenschaft – Schrift 176
Herausgegeben von Wolfgang Semar
Rade Jevdenic
Governance von Social-Media-Algorithmen im Digital Services Act
Analyse der Aufsicht und Regulation von
ML-basierten Empfehlungssystemen
Chur, 2024
ISSN 1660-945X

Churer Schriften zur Informationswissenschaft – Schrift 177
Herausgegeben von Wolfgang Semar
Ramona Kälin
Verantwortungs- & respektvoller Umgang im Metaverse
Eine Untersuchung, welche Rolle die Medienkompetenz spielt, wenn Jugendliche Hatespeech im Metaverse erfahren.
Chur, 2024
ISSN 1660-945X

Churer Schriften zur Informationswissenschaft – Schrift 178
Herausgegeben von Wolfgang Semar
Felicia Perrucci
Eine Erhebung des Status Quo der Therapiehun-de in Deutschschweizer Hochschulbibliotheken
Chur, 2024
ISSN 1660-945X

Churer Schriften zur Informationswissenschaft – Schrift 179
Herausgegeben von Wolfgang Semar
Sandra Morach
Webarchivierung im UZH Archiv
Erstellung einer Prozessbeschreibung und Erarbeitung von Empfehlungen für die Konzipierung eines Datenmodells sowie bezüglich der Wahl einer Preservation Planning-Strategie
Chur, 2024
ISSN 1660-945X

Über die Informationswissenschaft der Fachhochschule Graubünden

Die Informationswissenschaft ist in der Schweiz noch ein relativ junger Lehr- und Forschungsbereich. International weist diese Disziplin aber vor allem im anglo-ameri-kanischen Bereich eine jahrzehntelange Tradition auf. Die klassischen Bezeichnungen dort sind Information Science, Library Science oder Information Studies. Die Grundfragestellung der Informationswissenschaft liegt in der Betrachtung der Rolle und des Umgangs mit Information in allen ihren Ausprägungen und Medien sowohl in Wirtschaft und Gesellschaft. Die Informationswissenschaft wird in Chur integriert betrachtet.

Diese Sicht umfasst nicht nur die Teildisziplinen Bibliothekswissenschaft, Archiv-wissenschaft und Dokumentationswissenschaft. Auch neue Entwicklungen im Bereich Medienwirtschaft, Informations- und Wissensmanagement und Big Data werden gezielt aufgegriffen und im Lehr- und Forschungsprogramm berücksichtigt.

Der Studiengang Informationswissenschaft wird seit 1998 als Vollzeitstudiengang in Chur angeboten und seit 2002 als Teilzeit-Studiengang in Zürich. Seit 2010 rundet der Master of Science in Business Administration das Lehrangebot ab.

Der Arbeitsbereich Informationswissenschaft vereinigt Cluster von Forschungs-, Entwicklungs- und Dienstleistungspotenzialen in unterschiedlichen Kompetenzzentren:

- Information Management & Competitive Intelligence
- Collaborative Knowledge Management
- Information and Data Management
- Records Management
- Library Consulting
- Information Laboratory
- Digital Education

Diese Kompetenzzentren werden im Swiss Institute for Information Science (SII) zusammengefasst.

Impressum

Impressum

FHGR - Fachhochschule
Graubünden
Information Science
Pulvermühlestrasse 57
CH-7000 Chur

www.informationscience.ch
www.fhgr.ch
ISSN 1660-945X

Institutsleitung

Prof. Dr. Ingo Barkow
Telefon: +41 81 286 24 61
Email: ingo.barkow@fhgr.ch

Sekretariat

Telefon: +41 81 286 24 24
Fax: +41 81 286 24 00
Email: clarita.decurtins@fhgr.ch