Fairness im Kontext der Digitalisierung[1]

Was XING von Rawls und Kant lernen kann

Paula Becker und Julian Wagner

1. Einleitung

Definitionen und Verständnisse von dem, was als fair bezeichnet werden kann, werden von Menschen unterschiedlich aufgefasst. Beispielsweise charakterisieren einige anständiges Verhalten als fair, wohingegen andere Ehrlichkeit als Kriterium für Fairness betrachten. In der praktischen Anwendung wird schnell klar, dass zwar eine Vielzahl von ethischen Intuitionen besteht, diese aber offenlassen, was genau eine Handlung unfair macht. Dies wird dann am deutlichsten, wenn ethische Probleme nicht nur benannt, sondern auch aufgelöst werden sollen. Während beispielshaft Einigkeit über die moralische Intuition existiert, dass Amazons Recruiting-Tool nicht nur Männer einstellen, und der österreichische Arbeitsmarkt-Service Frauen nicht nur Teilzeitjobs vermitteln soll, gibt es keine Klarheit darüber, wie diese und andere Probleme im Detail aufgelöst werden sollten (vgl. Wilke 2018; Köver 2024). So wird die Frage nach einer allgemeingültigen Definition immer bedeutsamer, um ethischen Schwierigkeiten wirksam entgegentreten zu können Ein besonders aktuelles Beispiel dafür bietet die Ethik der Digitalisierung. Unternehmen und Konzerne spielen eine Schlüsselrolle darin, Fairness im Kontext der Digitalisierung zu leben und zu ermöglichen. Daher ist

1 Workshop auf der #CDRK24 Konferenz unter dem Titel „Fairness in KI Richtlinien: Durchsetzbar oder leeres Versprechen?“ – geleitet von Sebastian Riemann (XING – Part of New Work SE) und Leonhard Henke (CDR-Initiative), betreut von Julian Wagner (Universität Bayreuth) und Paula Becker (Universität Bayreuth). Der nachfolgende Beitrag ist im Rahmen eines Blockseminars „Ethik der Digitalisierung“ an der Universität Bayreuth entstanden und wurde durch den Workshop auf der #CDRK24 Konferenz inspiriert. Teile dieses Fachtexts wurden unter Verwendung generativer KI-Tools erstellt. Dabei wurde Chat GPT 4.0 zur Umformulierung von Textpassagen und zur Recherche genutzt. Keinerlei inhaltliche Aspekte wurden erstellt. Alle Ergebnisse wurden fachlich überprüft und bearbeitet.

es insbesondere im unternehmerischen Kontext wichtig, ein Grundverständnis von dem, was genau mit Fairness gemeint ist, zu beleuchten. Doch was genau bedeutet Fairness im Zuge der Digitalisierung und welche unternehmerischen Verantwortungen ergeben sich daraus? Um das Problem, vor dem wir stehen, greifbarer zu machen, hier ein Beispiel: Als professionelles Business-Netzwerk dient XING dazu, geschäftliche Kontakte zu pflegen und zu erstellen. Dabei verwendet XING Algorithmen, um Bewerber*innen und Betriebe miteinander zu verbinden. Jobangebote werden passenden Kandidat*innen vorgeschlagen und umgekehrt können Unternehmen Anzeigen für offene Stellen schalten. Die Fortschrittlichkeit dieser Algorithmen wird allerdings von einem Nachteil überschattet. Denn je nachdem, mit welchen Daten der Algorithmus gefüttert wurde, kann das Ergebnis des eingesetzten Programms dazu führen, dass bestimmte Gesellschaftsgruppen weniger berücksichtigt werden. Unternehmen wie XING beschäftigen sich bereits ausführlich mit der ethischen Verantwortung in Bezug der Digitalisierung und dem Einsatz von Künstlicher Intelligenz. Was in der Diskussion innerhalb der CDR-Konferenz 2024 allerdings auffiel, ist, das zwar über Fairness diskutiert wird, dabei aber keine branchenübergreifende Einigkeit darüber besteht, was wir eigentlich meinen, wenn wir über Fairness sprechen. Das ist auch der Grund, warum die einzelnen Definitionen von Fairness zwar nicht zufällig, aber doch als individuell und gewissermaßen subjektiv bezeichnet werden können. Dieser Beitrag hat daher das Ziel, der unternehmerischen Definition von Fairness von XING eine philosophische Definition gegenüberzustellen und zu untersuchen, was XING, trotz ihres Engagements in diesem Bereich, von Rawls und Kant über Fairness und deren Umsetzung in die Praxis lernen kann.

2. Definition von Fairness am Beispiel XING

2.1 Fairness bei XING

XING konzentriert sich als professionelles Business-Netzwerk mit 22.1 Millionen registrierten Nutzer*innen auf den deutschsprachigen Raum. Dabei ermöglicht es XING, Unternehmen und Privatpersonen in Kontakt zu treten und sich auf Stellenangebote zu bewerben. Warum passiert es nun, dass bestimmte Gesellschaftsgruppen wie beispielsweise Frauen oder Menschen mit dunkler Hautfarbe Unternehmen weniger häufig vorgeschlagen werden? Wie entstehen diese Ungleichheiten? Der Grund dafür liegt häufig in dem eingesetzten Algorithmus und insbesondere in den

Daten, mit denen dieser gefüttert wurde. Die Ursache hierfür findet sich in der sogenannten „Repräsentationsverzerrung", welche beispielsweise durch Stichprobenverzerrungen verursacht wird. Das bedeutet, wenn ein Algorithmus durch die verwendeten Daten gelernt hat, dass beispielsweise Frauen weniger häufig in Ingenieurberufen eingestellt werden, dann wird er mitunter aufhören, bestimmten Unternehmen Frauenprofile vorzuschlagen, obwohl diese genauso qualifiziert sind wie ihre männlichen Mitstreiter. Wie genau löst XING diese Ungleichheiten? Um zu verstehen, wie XING mit den bestehenden Problemen von „unfairen" Algorithmen umgeht, bedarf es zunächst einer Definition von dem, was XING als Fairness definiert: "Fairness bedeutet, dass KI-Systeme Entscheidungen treffen, die „keine" unberechtigten Vorurteile oder Diskriminierungen enthalten" (Reimann 2024a) Auch erklärt XING weiter:

> Fairness in KI bedeutet, Entscheidungen ohne Vorurteile oder Diskriminierung zu treffen [...]. Für uns bei XING steht im Fokus, dass Bewerberinnen und Bewerber mit unseren KI-gestützten Recruiting Tools allein auf Basis ihrer Kompetenzen und Fähigkeiten bewertet werden – unabhängig von Alter, Herkunft oder Geschlecht. Unser Ziel: eine möglichst objektive und diskriminierungsfreie Entscheidungsgrundlage. Damit fördern wir Chancengleichheit für alle bei der Jobsuche (Reimann 2024b).

An dieser Stelle ein kurzer Hinweis. Es ist nie eine künstliche Intelligenz an sich, die unfair entscheidet. Der Ursprung einer von uns als unfair wahrgenommenen Entscheidung liegt in den Daten, mit denen der Algorithmus gefüttert wurde. Das heißt, es sind Menschen, die (bewusst oder unbewusst) dafür verantwortlich sind, dass eine künstliche Intelligenz diese unfairen Vorschläge unterbreitet und beispielhaft Menschen mit dunkler Hautfarbe Unternehmen weniger häufig vorgeschlagen werden. Da eine Verzerrung der Daten nicht immer zu vermeiden ist, bleibt zu klären, wie genau XING den Anspruch auf Nichtverzerrung umsetzt. Um trotz einer Verzerrung der Daten in KI-Systemen, deren Umgang möglichst fair und transparent zu gestalten, hat sich XING nach eigenen Angaben dazu entschieden, auf die Idee von Margaret Mitchell et al. (2019) zurückzugreifen. Diese schlagen in ihrem Beitrag aus 2019 „Model Cards for Model Reporting", eine standardisierte Methode zur Dokumentation über KI-Modellen[2] vor. Dazu erklärt XING:

---

2 Als Modell werden wir einen fertigen Algorithmus bezeichnen, der mit verschiedenen Daten trainiert wurde.

> Um Fairness in unseren KI-Systemen aktiv voranzutreiben, bereiten wir die Model-Cards nach Mitchell et al. (2019) auf. Diese dokumentieren die relevanten Eigenschaften unserer Modelle, einschließlich realer und vermuteter Biases sowie Einschränkungen, und sorgen dafür, dass mit den Modellen unternehmensintern verantwortungsvoll umgegangen wird (Reimann 2024b).

2.2 Model Cards

Um XINGs Verständnis von Fairness besser greifen zu können, lohnt es sich, etwas genauer hinter das Konzept von Model Cards zu blicken. Was genau sind Model Cards und wieso helfen sie, Fairness im Unternehmen selbst und gegenüber den Nutzer*innen von XING zu leben? Die Gründerin von TrailML, Anna Spitznagel, sieht als grundlegende Idee von Model Cards, schwer begreifliche Themen leichter zugänglich zu machen. Model Cards sollen deshalb in erster Linie die Transparenz erhöhen und die Verantwortlichkeit bei der Nutzung von KI-Modellen verbessern. Das ist deshalb so wichtig, weil viele Dinge in der Entwicklung oder in der Arbeit mit KI-Modellen verloren gehen, da Informationen nicht gespeichert oder dokumentiert werden. Doch wie genau lösen Model Cards diese Probleme und was verbirgt sich hinter diesem Konzept?

Um die Transparenz in der Entwicklung und dem Umgang mit KI-Modellen zu verbessern, werden technische Details eines Modells vorgestellt und Risikoevaluationen durchgeführt. Das gelingt, indem anhand des Schemas einer Model Card genau dokumentiert wird, wie ein Modell getestet, entwickelt und eingesetzt wird (vgl. Mitchell et al. 2019). Die Motivation dieser Dokumentationsanleitung liegt darin begründet, dass maschinelle Lernmodelle häufig in sensiblen Bereichen wie Medizin, Strafrecht oder – wie in unserem Fall – Recruiting eingesetzt werden. Dabei mangelt es oftmals an Transparenz darüber, wie genau diese Modelle funktionieren. Eine Model Card setzt hier an und stellt, wie Abbildung 1 zu entnehmen ist, als strukturiertes Dokument Informationen über Zweck, Daten, Einschränkungen, empfohlene Anwendungen und Metriken eines KI-Modells zur Verfügung.

Neben den genannten Abschnitten schlagen Mitchell et al. außerdem eine Dokumentation über „Ethical Considerations" vor (vgl. Abbildung 1). In diesem Kapitel sollen Fragen nach der Nutzung von sensiblen Daten oder möglichen Risiken von der Nutzung des KI-Modells nachgegangen werden (vgl. ebd.: 225). Durch das Protokollieren von Biases und Einschränkungen von KI-Modellen ermöglicht XING damit ein besseres, unternehmensinternes Verständnis von dem, was ihre KI-

Modelle auszeichnet und welche Schwächen es im Umgang mit ihnen zu beachten gilt. Dieses Vorgehen kann als absolut fortschrittlich beschrieben werden, da Dokumentationen über die Entstehung und Nutzung von KI-Modellen nicht verpflichtend sind. XING ist sich dabei seiner ethischen Pflicht bewusst, die Transparenz ihrer Entwicklungen im eigenen Unternehmen zu fördern und zu kommunizieren (vgl. Abbildung 1).

Aufbau einer Model Card

Model Details	Basic Information about the model (e.g. Model date, Model version, licence, person or organziation developing the model, etc.)	*Evaluation Data*	Details on the dataset(s) used for the quantitative analyses in the card
Intended Use	Use cases that were envisioned during the development	*Training Data*	May not be possible to provide in pratice. When possible, this section should mirror evaluation data. If such detail is not possible, minimal allowable information should be provided here, such as details of the distribution over various factors in the training datasets
Factors	Factor could include demographic or phenotypic groups, environmental conditions, technical attributes, etc. (e.g. primary intended uses and users, out-of-scope use cases)	*Quantitative Analyses*	Unitary results, intersectional results
Metrics	Metrics should be chosen to relfect potential real-world impacts of the model (e.g. model performance measures, variation approaches)	*Ethical Considerations* *Caveats and Recommendations*	

ABBILDUNG 1: AUFBAU EINER MODEL CARD
(QUELLE: IN ANLEHNUNG AN MITCHEL ET AL. 2019: 222)

Trotz XINGs fortschrittlichem Verhalten, sich aktiv mit Fragen über Fairness in KI-Modellen zu beschäftigen und diese in ihrem Arbeitsalltag zu integrieren, scheint dem ethischen Problem von aufkommenden Ungleichheiten auf Netzwerken wie XING insbesondere zwei Herausforderung gegenüberzustehen. Zunächst ist es wichtig anzumerken, dass die von Mitchell et al. angesprochene fehlende Transparenz durch den Einsatz von Model Cards in einem wichtigen Aspekt nicht sichergestellt wird. Große, aus aber hunderttausend und mehr Parametern bestehende Modelle sind nämlich in ihrer genauen Funktionsweise für Entwickler*innen und Nutzer*innen undurchsichtig. Dies bedeutet, dass die einzelnen Entscheidungen eines derartigen Systems nicht nachvollziehbar sind. Zum Beispiel ist es nicht möglich nachzuvollziehen, welche Input-Variablen zu einem bestimmten Output geführt haben oder noch feiner aufgelöst, welche einzelnen Bausteine bzw. Parameter innerhalb des Netzwerkes ausschlaggebend für eine Entscheidung waren (vgl. von Eschenbach 2021). So kann zum Beispiel ein neuronales Netz, welches das Kreditausfallrisiko eines bestimmten Menschen bestimmt und damit für die Allokation von Krediten zuständig ist, einem Kreditantragsteller dessen Antrag abgelehnt wurde, keine Auskunft darüber geben, welcher individuelle Faktor oder Kombination von Faktoren (Input) für die Ablehnung (Output) verantwortlich war. Dies führt dazu, dass über die einzelne Entscheidung des Modells keine hundertprozentige Vorhersage getroffen werden kann.[3]

In Model Cards finden sich deshalb nur Auswertungen über das generelle durchschnittliche Entscheidungsverhalten des Systems über viele Einzelfälle. Diese beschriebene Undurchsichtigkeit hat auch eine moralische Komponente. Wenn eine mich betreffende Entscheidung von einem undurchsichtigen System getroffen wurde, welches mir seine Abwägung für meinen Einzelfall nicht darlegen kann, wie soll ich dann Handhabe gegenüber der Maschine haben und wie möchte der Betreiber die moralische Verantwortung für die besagte Entscheidung übernehmen? Die von uns fortlaufend als „Transparenz-Challenge" gekennzeichnete Herausforderung ist also insbesondere deshalb problematisch, da Nutzer*innen von XING kaum nachvollziehen können, wie genau ein Algorithmus funktioniert und warum einem Unternehmen nun bspw. ausgerechnet dieser

[3] Das hat auch als Konsequenz, dass die Folgen von KI-Modellen nicht direkt abzuschätzen sind. Siehe zu diesem Problem den Beitrag „Technologieentwicklung und Gerechtigkeit im Zeitalter der Digitalisierung. Die Diversity-Folgenabschätzung als Instrument zur Auflösung des Collingridge-Dilemmas" von Emily Breuer und Olivia Hankins (vgl. Breuer/Hankins 2025).

Bewerber und keine Bewerberin vorgeschlagen wird (vgl. Venkatasubramanian 2020). Diese Undurchsichtigkeit auf Seiten der Nutzer*innen hat eine erhebliche moralische Komponente, auf welche wir, gerüstet mit philosophischer Theorie, später eingehen werden.

Die „Transparenz-Challenge" ist eng verknüpft mit einer weiteren Problematik: Netzwerke wie XING liefern, trotz fairen Algorithmen, unfaire Ergebnisse. Wie genau lässt sich dieser Umstand erklären? Es ist empirisch belegt, dass zwar dem Problem von unfairen Algorithmen (also Verzerrungen in den Daten) entgegengewirkt werden kann, dies löst jedoch nicht den bestehenden Bias im Endergebnis (Outcome) auf. Was genau bedeutet das? Yulia Evsyukova et al. haben in ihrer Studie „LinkedOut? A Field Experiment on Discrimination in Job Network Formation" aus 2024 gezeigt, dass Menschen mit dunkler Hautfarbe in beruflichen Netzwerken diskriminiert werden. Es konnte gezeigt werden, dass die Wahrscheinlichkeit, dass Verbindungsanfragen von Profilen von Schwarzen Menschen akzeptiert werden, um 13 Prozent geringer ist als von nicht schwarzen Profilen. Der Grund dafür liegt demnach nicht in verzerrten Daten, sondern in dem diskriminierenden Verhalten von Individuen, die sich signifikant weniger häufig dazu entscheiden, sich mit Schwarzen Profilen zu verknüpfen.

Das Ergebnis dieser Studie deckt eine zweite grundsätzliche Ebene von ethischer Problematik auf. Denn es stellt sich die Frage nach der Verantwortung von Unternehmen wie XING, wenn das Endergebnis (Outcome), trotz fairem Algorithmus, durch eine ungleiche Behandlung von Individuen mit gleichen Qualifikationen charakterisiert wird. Diese Herausforderung wird fortlaufend als „Outcome-Challenge" bezeichnet. Inwieweit dabei diese Diskriminierung noch innerhalb der Verantwortung und des Pflichtbewusstseins des Netzwerkbetreibers liegt, ist abhängig von der Auslegung und Definition von Fairness.

An dieser Stelle scheint ein erster Hinweis in Bezug zu XINGs Definition von Fairness angebracht. Denn es lässt sich anmerken, dass das Fokussieren in XINGs Fairness Definition auf den Diskriminierungsbegriff problematisch ist. So zeigt auch schon die Entwicklung von einzelnen Anwendungen, dass es nicht möglich ist, ein Modell vollständig diskriminierungsfrei zu gestalten, wenn man das Wort im Wortsinn gebraucht, wovon ohne weitere Erklärung ausgegangen werden muss. Die Aufgabe eines Machine-Learning-Systems ist es, zu diskriminieren, in dem es einteilt, einordnet oder detektiert. Die Frage sollte also sein, welche Diskriminierung moralisch zulässig ist und welche nicht. So könnte das moralisch Verwerfliche einer Diskriminierung in der Inferenz von einer Subgruppe auf eine Einzelperson liegen. Diese Ansicht betont, dass Menschen als Individuen behandelt werden sollten, nicht nur als Teile einer Gruppe.

Das Problem dieser Argumentation besteht, ähnlich wie bei der Definition von Diskriminierung darin, dass moralische Verwerflichkeit zu pauschal zugeschrieben wird. Streng genommen wäre es beispielsweise eine unzulässige Diskriminierung, wenn Bewerber*innen aufgrund eines renommierten Hochschulabschlusses als besonders produktiv im Arbeitsalltag eingestuft werden (vgl. Binns 2018). Dies basiert auf der allgemeinen Annahme, dass das Individuum die Eigenschaften der gesamten Gruppe teilt und wäre aber auch dann diskriminierend, wenn diese Annahme durch empirische Daten gestützt und im konkreten Fall zutreffend wäre. Die Nutzung des Begriffes Diskriminierung erscheint uns deshalb sowohl im Wortsinn wie er von XING benutzt wird, als auch in einer etwas genaueren Ausarbeitung in Hinsicht unsere Challenges als nicht zielführend.

Nachdem nun einige bestehende Herausforderungen für XING beleuchtet wurden, stellt sich die Frage, wie mit diesen umgegangen werden sollte. Wie sollen wir von hier weiterdenken? Unser Vorschlag ist es, sich zwei stark verschränkten philosophischen Positionen zu widmen, die uns das Nachdenken über Fairness erleichtern und uns ermöglichen, klare Ableitungen und Handlungsempfehlungen auszusprechen. Dazu entwickeln wir in den kommenden Kapiteln spezifische und philosophisch gedachte Lösungen für die „Transparenz- sowie die Outcome-Challenge".

3. Challenges

3.1 Die Transparenz-Challenge – eine kantianische Perspektive

Als Transparenz-Challenge wurde das Problem bezeichnet, das Entwickler*innen und insbesondere Nutzer*innen kaum nachvollziehen können, wie Modelle und Algorithmen im Detail funktionieren und warum Entscheidungen so getroffen werde, wie sie getroffen werden. In diesem Kapitel wollen wir aufdröseln, warum diese Bestandsaufnahme insbesondere gegenüber den Nutzer*innen von XING eine moralische Komponente enthält, die es nach Kant nicht zu unterschätzen gilt. Dazu schauen wir uns zwei wichtige Konzepte aus Kants Philosophie an und erklären anhand dieser die moralische Wichtigkeit, der „Transparenz-Challenge" entgegenzuwirken.

Kant erklärt bereits in der Vorrede der Grundlegung zur Metaphysik der Sitten (1785/2016), dass er nach einem moralischen Gesetz suchen will, das mit absoluter Notwendigkeit gelten soll. Das heißt, Moral soll, analog zu rechtlichen Gesetzen, einer Gebundenheit unterliegen. Und es

gibt gute Nachrichten: Er findet ein solches Gesetz. Kant erklärt, dass moralisches Handeln gelingt, indem sich der Mensch an ein ganz bestimmtes Prinzip hält: den kategorischen Imperativ. Doch wie genau hilft dieser bei den bestehenden Schwierigkeiten, Fairness in der Digitalisierung zu leben? Zunächst eine kleine Begriffsklärung: Ein Imperativ beschreibt erst einmal ganz grundsätzlich eine objektiv geltende Handlungsnorm. Das bedeutet, ein Imperativ schreibt nicht einfach eine Handlungsregel vor, sondern sagt aus, dass diese tatsächlich notwendig ist. Imperative besitzen also einen nötigenden Charakter. Diese Eigenschaft ist für uns von besonderem Interesse, denn unsere Kritik besteht ja unter anderem darin, dass es eine solche Nötigung für bestehende Definitionen und Umsetzungen von Fairness bisher gerade nicht gibt.

Wir begegnen also mit dem kategorischen Imperativ einer Gesetzmäßigkeit, die unabhängig von subjektivem Empfinden ist. Was bedeutet diese Entdeckung für das Konzept Fairness und dessen Umsetzung in der Unternehmenswelt? Und was genau beschreibt nun dieses von Kant entdeckte moralische Prinzip? Es gibt unterschiedliche Formen des kategorischen Imperativs, doch die allgemeinste Formel lautet: „[H]andle nur nach derjenigen Maxime, durch die du zugleich wollen kannst, daß sie ein allgemeines Gesetz werde“ (Kant 1785/2016: Abschnitt 421, Zeile 7 f.). Mit diesem berühmten Ausdruck meint Kant, dass der Mensch sich bei moralischen Überlegungen oder Unsicherheiten die Frage stellen soll, ob das Prinzip, nach dem er handelt, auch dann noch gewollt werden kann, wenn es universalisiert wird. Wenn also jeder Mensch unter allen Umständen nach eben diesem Prinzip handelt. Dabei unterziehen wir einem subjektiven Prinzip zu handeln (einer Maxime) einer Art Test. Kann der Mensch vernünftigerweise wollen, dass dieses Prinzip universalisiert wird, dann hat es den Test bestanden. Besteht die Maxime den Test nicht, dann gilt sie als unmoralisch und darf nicht ausgeführt werden. Das ist immer dann der Fall, wenn die Maxime in eine von zwei Arten von Widersprüchen gerät. Einmal darf die universalisierte Maxime keinen Widerspruch begrifflicher Art hervorrufen. Das bedeutet, das subjektive Prinzip zu Handeln darf in keinen logischen Widerspruch führen. Am Beispiel der Lüge wird das besonders deutlich. Denn wenn es meine Maxime ist, in einer Notsituation zu lügen und meinem Gegenüber zu versprechen, ihm das geliehene Geld zurückzahlen, ohne dies zu beabsichtigen, dann wird das Lügen an sich bedeutungslos. Das liegt daran, dass, wenn jeder diese Maxime verfolgen würde, die Lüge als solche nicht mehr funktioniert (vgl. ebd.: Abschnitt 422). Die Maxime hat den Test ebenso wenig bestanden, wenn dieselbe einen Widerspruch im Willen selbst herbeiführt. Hierzu ein weiteres Beispiel: Wenn ich mich dazu entscheide, meine Talente nicht zu entwickeln

und mich nicht darum bemühe, mich weiterzuentwickeln, dann kann dieses Prinzip zu Handeln zwar ohne logischen Widerspruch gedacht werden, aber Kant spricht davon, dass ich es "unmöglich wollen" (ebd.: Abschnitt 423, Zeile 28 f.) kann.

Um nun die Tragweite der „Transparenz-Challenge" zu verdeutlichen, können wir den Test auf eine beispielhaft gedachte Maxime von XING anwenden. Ein mögliches Handlungsprinzip könnte wie folgt aussehen: Solange der Algorithmus grundsätzlich faire Ergebnisse liefert, muss meinen Kund*innen nicht deutlich werden, wie diese zustande kommen. Nun fragt sich XING, kann ich wollen, dass meine Maxime ein allgemeines Gesetz wird – das also jeder Mensch nach diesem Prinzip handelt? Ist es vernünftig anzunehmen, dass faires Verhalten gelingt und umsetzbar wird, wenn diejenigen, die es betrifft, nicht darüber informiert werden, wie diese Ergebnisse zustande kommen? Wenn also Nutzer*innen gar nicht selbst darüber urteilen können, ob es fair ist, dass sie beispielshaft einem erhofften Unternehmen nicht angezeigt wurden? Es lässt sich vermuten, dass XING zu dem Schluss kommt, dass dies nicht vernünftigerweise gewollt werden kann. Denn ein solches Prinzip, zu handeln, würde für XING bedeuten, dass dem Unternehmen ebenfalls kein vollständiges Wissen über Entscheidungen zusteht, die es selbst betrifft. Das heißt, auch XING würde unter intransparenten Entscheidungen Anderer leiden. Die Undurchsichtigkeit auf der Seite der Kund*innen von XING besteht darin, dass nicht genau ersichtlich wird, wie bestimmte Verknüpfungen zustande gekommen sind. Eine mögliche Undurchsichtigkeit auf der Seite von den Mitarbeiter*innen von XING kann wiederum darin bestehen, dass XING mit Partnerfirmen zusammenarbeitet, die ihre Algorithmen nach anderen Prinzipien und ethischen Leitfäden konstruiert, wie XING das tut. Wenn keine Maxime von Transparenz besteht, dann leidet XING, in diesem Fall nun als Kunde oder Geschäftspartner, selbst unter seinem eigenen Prinzip zu Handeln. Wir begegnen einem Widerspruch im Willen selbst, da dies von XING nicht vernünftigerweise gewollt werden kann; die Maxime hat den Test also nicht bestanden.

Um zu verstehen, warum die „Transparenz-Challenge" ein echtes moralisches Problem darstellt, darf ein weiteres Konzept aus Kants Philosophie nicht fehlen: die Menschenwürde und damit einhergehend die sittliche Autonomie des Menschen (vgl. Willaschek et al. 2015: 2693). Als Würde wird bei Kant ein absoluter und nicht gegenrechenbarer Wert des Menschen bezeichnet. Diese Würde ist dem Menschen dabei praktisch zuzurechnen und weder dinglich konditioniert noch anderweitig hergeleitet. Sie ist dem Menschen qua seines Menschseins zuzuschreiben. Dabei basiert sie auf dem Gedanken der Autonomie. Das bedeutet, der Ursprung der Menschenwürde

liegt der menschlichen Fähigkeit zugrunde, sich selbst zum Handeln zu bestimmen. In dieser Funktion wird Autonomie mit Freiheit gleichgesetzt. Freiheit ist in diesem Sinne allerdings nicht so zu verstehen, dass der Mensch an kein Gesetz gebunden ist. Es ist vielmehr so, dass der Mensch an Gesetze gebunden ist, die er sich in gewisser Weise selbst auferlegt hat (vgl. Johnson/Cureton 2024). Nun scheint ein eingeschränktes Verständnis oder zumindest der Zugang zu Informationen von künstlicher Intelligenz dieser Freiheit zu widersprechen. Denn wenn nicht gewusst oder verstanden wird, wie bestimmte Entscheidungen oder Vorschläge zustande kommen, dann kann zum einen keine passende Maxime aufgestellt werden und zum anderen fehlt es den Nutzer*innen an der Möglichkeit, sich im kantischen Sinne so frei wie möglich zu entfalten und damit selbst auferlegte Gesetze adäquat zu verfolgen. Ein solches Gesetz könnte beispielshaft sein, sich in der Arbeitswelt für diejenigen Unternehmen zu entscheiden, die die eigenen Werten am besten widerspiegeln. Undurchsichtige Algorithmen können dieses Prinzip erschweren oder verhindern, wenn nicht deutlich wird, aus welchen Gründen bestimmte Unternehmen vorgeschlagen werden. Dies widerspricht aber der Autonomie des Menschen und damit zusammengehörig auch der Menschenwürde.

Das es aus moralischer Sicht also unzulässig ist, die „Transparenz-Challenge" einfach zu akzeptieren, wurde durch den Begriff der Menschenwürde und die Anwendung des kategorischen Imperativs deutlich. Denn, wenn jeder Mensch nach der beispielhaft genannten Maxime von XING handeln würde, würde es kaum gelingen, strukturellen Ungleichheiten und unfairen Ergebnissen auf Plattformen wie XING entgegenzuwirken. Auch betroffen ist dabei die Menschenwürde, insofern sich der Mensch nicht mehr adäquat zum Handeln bestimmen kann, wenn er die Ursprünge der ihn betreffenden Entscheidungen nicht nachvollziehen kann. Fragt man also Kant, sollten die Nutzer*innen nicht im Unklaren darüber bleiben, wie ein verwendeter Algorithmus Entscheidungen über Profilvorschläge etc. trifft.

3.2 Die Outcome-Challenge – eine rawlsianische Perspektive

Während sich Immanuel Kant mit der Frage von Fairness auf der Ebene des Individuums beschäftigt, stellt John Rawls die Frage nach einer fairen Gesellschaft. Was würde nun passieren, wenn sich Kantisches Denken nicht auf die Generalisierung (in Form einer Maxime) einer einzelnen Entscheidung bezieht, sondern auf die Organisation einer gesamten Gesellschaft? Diese

Frage versucht John Rawls in seinem Werk „Justice as Fairness“ zu beantworten. Um dies zu tun, vereint Rawls wesentliche Aspekte von Kants Philosophie mit seiner eigenen Vertragstheorie.

Für unsere Zwecke können wir Rawls kontraktualistische Basis auf die Idee zurückführen, dass die Einführung von gesellschaftlichen Institutionen sowie gewisse moralische Grundregeln in Form eines Gesellschaftsvertrages festgelegt werden können. Diese Grundregeln entstammen aus der Rationalität der dem Vertrag zustimmenden Personen. Das äußert sich darin, dass die abgeschlossenen Regeln dem Eigeninteresse der Einzelpersonen dienen. So argumentierte zum Beispiel schon John Locke, dass Menschen von Vernunft geleitet sind und natürliche Rechte auf Leben, Freiheit und Eigentum besitzen (vgl. Uzgalis 2024). Um diese Rechte besser zu schützen, treten sie in einen Vertrag ein, bei dem sie einer Regierung zustimmen, die auf ihrem Konsens beruht und deren Hauptaufgabe der Schutz dieser Rechte ist. Wie ein solcher Vertrag konstruiert werden soll, hängt maßgeblich von den Annahmen über die Welt und dem Menschen in dieser ab. Rawls innovativer Gedanke ist es nun, Menschen, die über die Ordnung der Gesellschaft entscheiden sollen, vor ein schon erwähntes Gedankenexperiment zu stellen.

Man stelle sich vor, man befände sich in einer Situation, in der man entscheiden soll, welche Prinzipien die Grundstruktur sowie die Institutionen einer Gesellschaft bestimmen sollen. Allerdings gibt es eine entscheidende Einschränkung: Man weiß nichts über seine eigene Stellung in dieser Gesellschaft. Man weiß nicht, ob man reich oder arm, gesund oder krank, gebildet oder ungebildet, Mann oder Frau sein wird. man kennt weder seine ethnische Zugehörigkeit noch seine Religion oder Talente. Diese Unwissenheit bezeichnet Rawls als „Schleier des Nichtwissens“. Hinter diesem Schleier des Nichtwissens haben Sie nun keine Idee davon, wie sich Ihre Entscheidungen auf Ihren persönlichen Einzelfall auswirken werden, sondern nur davon, wie sich dies auf alle möglichen Positionen auswirken könnte (vgl. Rawls 1999: 118–123).

Rawls nimmt nun an, dass Menschen hinter diesem Schleier des Nichtwissens den gleichen moralischen Ansprüchen genügen müssen, wie sie schon Kant formulierte. Um uns zu erinnern: Kants Konstruktion des kategorischen Imperativs stellt einen Vertrag des Menschen mit sich selbst dar, zu dessen Befolgung uns unsere Rationalität zwingt. Rawls hebt dies nun auf die Ebene des Gesellschaftsvertrages, welcher in den beschriebenen besonderen Umständen geschlossen wird. Welche Prinzipien ergeben sich nun daraus? Rawls formuliert zwei Prinzipien der Gerechtigkeit, nach welchen eine Gesellschaft geformt werden sollte:

> FIRST PRINCIPLE: Each person is to have an equal right to the most extensive total system of equal basic liberties compatible with a similar system of liberty for all.
>
> SECOND PRINCIPLE: Social and economic inequalities are to be arranged so that they are both: (a) to the greatest benefit of the least advantaged, consistent with the just savings principle, and (b) attached to offices and positions open to all under conditions of fair equality of opportunity (ebd.: 53).

Diese Prinzipien sind, wie schon erwähnt, ausgerichtet auf die Einrichtung der grundlegenden Institutionen einer Gesellschaft. Wie genau sollen wir nun mit diesen Prinzipien umgehen? Als erstes ist festzustellen, dass für Rawls der Gerechtigkeitsbegriff stark geknüpft ist an den Begriff der Fairness, welcher zentraler Gegenstand unseres Beitrages ist. Auch wenn Fairness nicht explizit in den beiden hier gezeigten Prinzipien erwähnt wird, sieht Rawls diese Prinzipien als zwingend, damit eine Institution fair sein kann. Weiter und expliziter formuliert er ein Prinzip der Fairness, welches auf der Ebene des Individuums seine Gültigkeit findet.

Um nun von der theoretischen philosophischen Betrachtung wieder zurück zu unserem Anwendungsfall zu kommen, ist eine Einordnung vonnöten. Wie kann XING von Rawls Prinzipien der Gerechtigkeit betroffen sein, wenn das Unternehmen keine klassische soziale Institution darstellt? Diese Frage hängt stark damit zusammen, wie wir den Begriff der Institution deuten. Auch wenn XING keine klassische Institution ist, so ist XING doch als Firma in relevanter Weise an der Verteilung von sozial bedeutsamen Gütern und auch Positionen (welche zwar nicht politische Ämter beinhalten, aber trotzdem soziale Stellungen darstellen) beteiligt (vgl. ebd.: 78–81). In dieser Weise ist das Handeln von XING gesamtgesellschaftlich relevant und hat somit Auswirkungen auf die Funktionsweise dieser Gesellschaft. Zurück unter dem Schleier des Nichtwissens ist uns zum Beispiel rationalerweise daran gelegen, dass Frauen dieselben Chancen haben, Netzwerke auf Plattformen aufzubauen wie Männer.

Wenn wir nun nach dieser Einordnung den Fall des diskriminierenden Outcomes in Bezug auf den Netzwerkaufbau von Schwarzen Profilen betrachten, sticht uns sofort Rawls zweites Prinzip der Gerechtigkeit ins Auge, welches sich mit sozialer und ökonomischer Ungleichheit auseinandersetzt. Für XINGs Handeln sind in diesem Fall beide Bausteine des Prinzips relevant. Im ersten Schritt lässt sich XINGs Wirtschaften so einordnen, dass sie durch die Zusammenführung

von Arbeitgeber*innen und Arbeitnehmer*innen einen relevanten Einfluss auf die Verteilung von Ämtern und Positionen haben. Damit unterliegen sie auf jeden Fall der Anforderung, Gleichheit der Bedingungen herzustellen. Im zweiten Schritt werden auch klare Konditionen für etwaige Ungleichbehandlungen gegeben: Ungleichbehandlung ist nur dann gerechtfertigt, wenn sie der Besserstellung der Schlechtgestelltesten dient. Die Betrachtung von ethischen Abwägungen aus der Sicht von Rawls bedeutet für XING den Perspektivwechsel in die Richtung, dass XING sich die Auswirkungen des Handelns auf die gesamte Gesellschaft vor Auge führen muss.

Durch Rawls lässt sich dies aber im zweiten Schritt schon konkret an der Einzelentscheidung (z. B. der Funktionsweise eines einzelnen Modells) festmachen. Es ist also im rawlsianischen Sinne nicht genug, dass das KI-System selbst fair entscheidet, wenn die Entscheidung keine Chancengleichheit gewährleistet. Mit Rawls gedacht, fällt also diesem, außerhalb der Entscheidungsmacht des Unternehmens liegenden Aspekt der Ungleichheit genauso viel moralisches Gewicht zu wie solchen, die dem eingesetzten Algorithmus zuzuschreiben sind. Inwieweit ist nun XING in unserem Beispiel verantwortlich für die Diskriminierung, welche von den Nutzer*innen ihrer Plattform ausgeht?

Rawls verdeutlicht, dass XING zwar keine klassische soziale Institution darstellt, jedoch aber den entsprechenden moralischen Grundsätzen gerecht werden muss. Damit einher geht auch, dass solche Institutionen moralische Ansprüche nicht nur vertreten sollen, sondern auch in ihrem Zuständigkeitsbereich für ihre Einhaltung verantwortlich sind. Dafür ist in Rawls Prinzipien auch explizit eine Kondition für Ungleichbehandlung vorgesehen, nämlich wenn sie den Schlechtgestelltesten dient. Hierzu soll erst einmal nur gesagt werden, dass es also zumindest eingeschränkt nicht gelten kann, dass XING bloß unbeteiligter und passiver Zuschauer bei „unfairen" Marktergebnissen ist, selbst wenn die eigene Vorgehensweise fair (z.B nichtdiskriminierend) ist.

Wie die „Outcome"- sowie die „Transparenz-Challenge" unserer Ansicht nach in relevanter Weise Auswirkungen auf das Handeln von XING haben sollte und welche praktischen Handlungsempfehlungen sich aus den gewonnen philosophischen Einblicken ableiten lassen, werden wir im Folgenden betrachten.

4. Was kann XING von Rawls und Kant lernen?

4.1 Handlungsempfehlungen

Transparenz praktizieren – nach dem Vorbild des kategorischen Imperativs

Um die von Kant geforderte Transparenz im Falle XINGs herzustellen, gilt es zwei Probleme zu lösen, welche auf die jeweiligen Formen der Intransparenz abzielen. Diese sind zum einen die Intransparenz gegenüber den Nutzer*innen und zum anderen die Intransparenz gegenüber den Entwickler*innen. Vorweg ist anzumerken, dass die Intransparenz von Modellen in Bezug auf beide Problematiken nur in Teilen gelöst werden kann, da ein vollständig erklärbares Modell noch außerhalb des technisch Möglichen liegt. Somit bleibt Kants Kritik gegenüber der Unmöglichkeit der Formulierung eines Imperativs beim Einsatz von Deep-Learning Modellen bestehen. Die einzige Ableitung aus diesem Fakt kann sein, erklärbare und interpretierbare Modelle zu verwenden, welche einer einsehbaren Logik folgen und damit deterministisch handeln.[4] Es gibt jedoch auch gute Nachrichten, denn mittlerweile existieren etablierte Möglichkeiten, einzelne Teilbereiche eines Modells zu erklären. Zwei explizite Methoden möchten wir beispielhaft hervorheben.

Es ist möglich, mit einer dem Entwicklungsprozess angeschlossenen Methode die sogenannte „Feature Importance" zu berechnen. Diese liefert Informationen darüber, welche Inputvariable welchen prozentualen Einfluss auf ein bestimmtes Endergebnis hat. So könnte zum Beispiel im Falle des schon betrachteten Modells, welches Kunden Kredite anhand eines Kreditausfallrisikos zuteilt, diejenigen Variablen, welche zu einer bestimmten Entscheidung geführt haben, aufgeschlüsselt und mit einer Gewichtung versehen werden (vgl. Karimi 2020).

Eine weitere Möglichkeit, in diesem Fall Transparenz herzustellen und die Handlungsfähigkeit zu gewährleisten, ist die Kommunikation mit den Nutzer*innen darüber, wie diese ihre Handlungsweisen optimalerweise anpassen können, um das Ergebnis bei einer weiteren Evaluation im gewünschten Sinne zu beeinflussen. Glücklicherweise gibt es für diese Art der Erklärung etablierte Methoden, namentlich die der „Counterfactual Explanations". Im Kontext der Kreditvergabe bedeutet dies beispielsweise, dass ein Kunde, dessen Antrag abgelehnt wurde, nicht nur erfährt, dass

4 Eine solche Ansicht vertritt zum Beispiel Cynthia Rudin, welche für den Verzicht auf undurchsichtige Modelle bei Hochrisiko-Anwendungen plädiert (vgl. Rudin 2019). Etwaige Performance Einbußen müssen laut Rudin dann entweder durch Forschung mitigiert oder akzeptiert werden.

sein Einkommen und seine Kreditwürdigkeit wesentliche Einflussfaktoren waren, sondern auch, welche konkreten Anpassungen zu einer Bewilligung geführt hätten. Eine kontrafaktische Erklärung könnte lauten: „Hätte Ihr monatliches Einkommen 500 Euro höher gelegen oder Ihr Kredit-Score um 20 Punkte besser abgeschnitten, wäre Ihr Antrag genehmigt worden". Dadurch wird nicht nur Transparenz geschaffen, sondern es entstehen auch klare Handlungsoptionen für den Antragsteller.

Dieses Beispiel zeigt, dass die genaue Gestaltung einer Erklärung individuell auf Einzelfälle angepasst werden muss, um die Sinnhaftigkeit der Information für die jeweiligen Nutzer*innen zu gewährleisten. Dieser Umstand und die Notwendigkeit, für bestimmte Anwendungen speziell entwickelte Lösungen einzusetzen macht es herausfordernd, Erklärbarkeit innerhalb XINGs Plattform zu integrieren. Während wir kein völlig ausgearbeitetes Konzept liefern können, wie XING in dem von uns angesprochenen Beispiel die Erklärbarkeit ihrer Systeme herstellen könnte, wollen wir zumindest eine Handlungsempfehlung für die generelle Herangehensweise an die Entwicklung und den Einsatz solcher Systeme liefern. Inspiration dabei könnte zum Beispiel das Framework für Erklärbarkeit von Markus Langer et al. liefern. Dieses stellt in das Zentrum der Entwicklung und des Einsatzes von Erklärbarkeitstechniken die betreffende Zielgruppe (vgl. Langer et al. 2021).

Wenn Techniken der erklärbaren Künstlichen Intelligenz (XAI) eingesetzt werden sollen, müssen nach Langer et al. verschiedene Aspekte beachtet werden. Zunächst ist es entscheidend, die relevanten Stakeholder zu identifizieren, da unterschiedliche Gruppen, etwa Nutzer*innen, Entwickler*innen oder Entscheidungsträger spezifische Anforderungen an die Erklärbarkeit haben. Diese Desiderata[5] müssen mit dem Ziel der Erklärung abgeglichen werden, um sicherzustellen, dass die bereitgestellten Informationen tatsächlich den jeweiligen Bedürfnissen entsprechen, sei es zur Erhöhung des Vertrauens, zur Handlungsermächtigung der Nutzer*innen, zur Einhaltung regulatorischer Vorgaben oder zur Verbesserung der Nutzbarkeit im Generellen.[6]

---

5 Das Wort Desiderat kommt aus dem Lateinischen und bedeutet so viel wie „Ersehntes". Ein Desiderat bezeichnet also ein Wunschobjekt.

6 Hier ist anzumerken, dass nicht nur ethische Aspekte, welche wir hier betonen, für den Einsatz erklärbarer KI sprechen, sondern auch Aspekte wie gesteigertes Vertrauen auf der Seite der Nutzer*innen, Effizienz in der Entwicklung sowie gesteigerte Kundenzufriedenheit durch solche Tools erreicht werden können.

Darauf aufbauend gilt es, den passenden Erklärungsansatz zu wählen, welcher auf die Bedürfnisse der bestimmten Stakeholder eingeht. Außerdem ist ein kontinuierlicher Feedback-Loop essenziell: Die eingesetzten Methoden müssen regelmäßig evaluiert, auf ihre Wirksamkeit geprüft und iterativ[7] verbessert werden, um den sich wandelnden Anforderungen der Stakeholder anzupassen. Es könnte eventuell nicht ganz vorhergesagt werden, welche Zielgruppe aus was für einer genauen Erklärung welchen Nutzen zieht. Zum Beispiel kommen für manche Anwendungen sowohl Feature Importance als auch Counterfactual Explanations in Frage und es muss dann iterativ eruiert werden, welche der beiden Methoden am besten funktioniert.

Entwickler*innen benötigen vor allem technisch aufgeschlüsselte Informationen zu den verschiedensten Spezifikationen eines Modells, um Modelle zu optimieren. Diese Erklärung kann und sollte sprachlich sowie inhaltlich an diese technische Expertise angepasst werden. Endnutzer*innen hingegen erwarten verständliche und nachvollziehbare Erklärungen für KI-Entscheidungen, die ihnen ermöglichen, daraus weitere Schritte abzuleiten oder entsprechend zu handeln. Diese Erklärungen müssen dann zum Beispiel in einer viel weniger technischen Sprache verfasst werden. Um dem gerecht zu werden, ist es wichtig, die Zielgruppe mit ihren Voraussetzungen (Wissen, Zeit-Ressourcen, Ansprüchen etc.) genau zu kennen und die Anwendung auf diese zuzuschneiden. Des Weiteren können Erklärungen auch auf Nachfrage der Nutzer*innen angepasst werden und damit beliebig ausführlich ausfallen, wenn dies gewünscht ist. Entwickler*innen hingegen benötigen detaillierte technische Informationen zur Optimierung des Empfehlungssystems, welche sich in Form, Sprache und Inhalt drastisch von denen der Nutzer*innen unterscheiden.

Ein weiterer Aspekt der Intransparenz betrifft die Intransparenz im Entwicklungsprozess. Einzelne Entscheidungen und Abwägungen werden iterativ im Entwicklungsprozess getroffen, jedoch nicht dokumentiert. Falls eine Dokumentation stattfindet, wie in kleinem Maße in den von XING genutzten Model Cards sind diese nur für einen engen Kreis an Entwickler*innen verständlich und dadurch auch nur für diese nutzbar. Model Cards dokumentieren jedoch auch nur den Endstand eines Modells und nicht den gesamten Entwicklungsprozess und machen es so nicht möglich, einzelne Entscheidungen innerhalb des Prozesses nachzuvollziehen. Eine solche Nachverfolgungsmöglichkeit ist in gleichem Maße notwendig, damit moralische Verantwortung für das Endprodukt und dessen Entscheidungen übernommen werden kann, genauso wie die

7 Iterativ beschreibt, dass der Prozess immer aufs Neue wiederholt wird.

Transparenz des Modells selbst. Entwicklungen in diese Richtung liefert zum Beispiel die Anwendungen des bereits erwähnten Münchner Startups TrailML. TrailML hat eine Anwendung entwickelt, die innerhalb der Entwicklungsumgebung wichtige Parameter sowie Hypothesen und Ergebnisse automatisch dokumentiert. So wird es Entwickler*innen einfach gemacht, den Entwicklungsprozess zu dokumentieren. Außerdem ist die Anwendung darauf ausgelegt, nach EU AI Act[8] konforme Berichterstattung zu leisten.

Gerechtigkeit gestalten – nach dem Vorbild des Differenzprinzips

Die für uns relevanteste Erkenntnis aus Rawls Überlegungen ist die Weitung des Blickes von Fairness innerhalb einzelner von XING eingesetzten Modellen, hin zu einer Betrachtung von Fairness im Gesamtergebnis. Entscheidend ist also die Chancengleichheit bei der Vermittlung von Arbeitnehmer*innen und Arbeitgeber*innen, welche sich im Endergebnis äußert. XING als Betreiber des Netzwerkes sollte nach Rawls als Institution entsprechend ihres Handlungsspielraums Verantwortung übernehmen. Im ersten Schritt lässt sich daraus ableiten, dass die Evaluation der von XING eingesetzten Algorithmen in Bezug auf den Fairness-Begriff auf die Handlungen der Nutzer*innen erweitert werden sollte. Dafür sind experimentelle Studien wie die von uns erwähnte Studie der Universität Mannheim entweder selbst durchzuführen oder durch entsprechende Wissenschaftler*innen durchführen zu lassen.

Wenn ein Missstand wie in der von uns angeführten Studie (vgl. Evsyukova 2025) auch im Netzwerk von XING besteht (wovon stark auszugehen ist), sollten Maßnahmen zur Mitigation[9] ergriffen werden. Nur wie sollten diese aussehen? Anzumerken ist hier der Handlungsspielraum, über den XING verfügt. Es ist nicht von XING zu erwarten, die in der Gesellschaft verankerten Vorurteile an der Wurzel des Problems zu beheben. Es ist jedoch innerhalb des Handlungsspielraums der Firma durchaus möglich, mit einem gewissen Maße an positiver Diskriminierung die Sichtbarkeit von Schwarz gelesenen Profile zu erhöhen, um Chancengleichheit herzustellen. Der Algorithmus würde also anhand einer anderen Variabel evaluiert werden, nämlich der gleichverteilten Rate von Netzwerkaufbauten für verschiedene Gruppen von Menschen. Solch eine Evaluation

8 Die Europäische Union hat neue Rechtsvorschriften zur künstlichen Intelligenz erlassen: den EU AI Act. Er schafft die Grundlagen für die Regulierung von KI in der EU.

9 Mitigation bedeutet „Abschwächung" oder „Milderung". Es geht um Maßnahmen, die die Ursachen oder Folgen eines Problems abschwächen.

durchzuführen, sollte den Entwicklern von XING höchstwahrscheinlich leichter fallen dadurch, dass sie besseren Zugriff auf die Daten des Netzwerkes haben. In das gesamte System wird dadurch jedoch auch erhebliche Komplexität eingeführt. Diese kommt daher, dass das Handeln der Nutzer*innen nun auch Auswirkungen auf die Rate der positiven Diskriminierung hat. Eine Anpassung im Algorithmus bedarf also eine hoch frequentierte Evaluation. Diesen Prozess könnten jedoch auch informierte Nutzer*innen erleichtern. Das wäre ein weiterer Vorteil der von uns im vorherigen Abschnitt erläuterten Maßnahmen zur Verringerung der Intransparenz.[10] Informierte Nutzer*innen können einsehen, ob eine Entscheidung über sie fair getroffen wurde und falls sie einen Missstand vermuten, diesen direkt effizient und genau an XING kommunizieren. Nur wenn das von XING eingesetzte System so angepasst und evaluiert wird, kann sichergestellt werden, dass es einer robusten Definition von Fairness, welche wir versucht haben dazustellen, gerecht wird.

5. Fazit

Trotz XINGs fortschrittlichem Verhalten und der Auseinandersetzung mit Fairness im Kontext der Digitalisierung konnte gezeigt werden, dass insbesondere zwei Herausforderungen bestehen bleiben: die „Transparenz"- sowie die „Outcome-Challenge". Durch die Auseinandersetzung mit den eng verknüpften Philosophien von Immanuel Kant und John Rawls wurde deutlich, dass diesen Challenges ein bedeutsames philosophisches Fundament zugrunde liegt. Um beiden Challenges entgegenzuwirken, wurden anschließend konkrete Handlungsempfehlungen ausgearbeitet, die als Lösungs- und praktische Verbesserungsvorschläge dienen können. Die „Transparenz-Challenge" kann so insbesondere dadurch gelöst oder verbessert werden, indem die Methode der „Feature Importance" oder der „Counterfactual Explanations" angewendet wird. Um die Challenge nach innen zu lösen, wurde außerdem auf das Unternehmen TrailML aufmerksam gemacht. Durch die Anwendung dieser Methoden kann es XING gelingen, mehr Transparenz gegenüber seinen Nutzer*innen zu ermöglichen. Als Lösungsvorschlag für die „Outcome-Challenge" wurde der Einsatz von Fairness-Parametern, welche über das System hinausgehen, motiviert. Namentlich die vergleichende Betrachtung des Netzwerkaufbaus von verschiedenen Nutzer*innengruppen. Mit diesem Wissen kann dann durch Rawls Argumentation positive Diskriminierung eingesetzt werden, um ein wirklich faires Endergebnis zu erzielen.

---

10 Einordnung: Nicht gegen alles kann und sollte positiv diskriminiert werden.

Binns, R. (2018): Fairness in Machine Learning: Lessons from Political Philosophy. Proceedings of the 1st Conference on Fairness, Accountability and Transparency, in: Proceedings of Machine Learning Research, Jg. 81, 149–159.

Breuer, E. / Hankins, O. (2025): Technologieentwicklung und Gerechtigkeit im Zeitalter der Digitalisierung. Die Diversity-Folgenabschätzung als Instrumetn zur Auflösung des Collingride-Dilemmas, in: Brink, A. (Hrsg.): Fairness im Zeitalter von KI, Baden-Baden: Nomos [im Erscheinen].

Evsyukova, Y. / Rusche, F. / Mill, W. (2025): LinkedOut? A Field Experiment on Discrimination in Job Network Formation, in: The Quarterly Journal of Economics, Jg. 140 / Nr. 1, 283–334, DOI: 10.1093/qje/qjae035.

Freeman, S. (Hrsg.) (2003): The Cambridge Companion to Rawls, Cambridge: Cambridge University Press.

Johnson, R. / Cureton A., (2024): Kant's Moral Philosophy, in: Zalta E. N. / Nodelman U. (Hrsg.): The Stanford Encyclopedia of Philosophy, URL: https://plato.stanford.edu/archives/fall2024/entries/kant-moral/ (aufgerufen am: 11/03/2025).

Kant, I. (1785/2016): Grundlegung zur Metaphysik der Sitten, Riga: J. F. Hartknoch.

Köver, C. (2024): Diskriminierung: AMS erntet Hohn mit neuem KI-Chatbot, in NETPOLITIK.ORG, URL: https://netzpolitik.org/2024/diskriminierung-ams-erntet-hohn-mit-neuem-ki-chatbot/#netzpolitik-pw (aufgerufen am: 25/03/2025).

Karimi, A. H. / Barthe, G. / Schölkopf, B. / Valera, I. (2020): A Survey of Algorithmic Recourse: Definitions, Formulations, Solutions, and Prospects, DOI: 10.48550/arXiv.2010.04050.

Langer, M. / Oster, D. / Speith, T. / Hermanns, H. / Kästner, L. / Schmidt, E. / Sesing, A. / Baum, K. (2021): What do we want from Explainable Artificial Intelligence (XAI)? – A Stakeholder Perspective on XAI and a Conceptual Model Guiding Interdisciplinary XAI Research, in: Artificial intelligence, Jg. 296, 103473.

Mitchell, M. / Wu, S. / Zaldivar, A. / Barnes, P. / Vasserman, L. / Hutchinson, B. / Spitzer, E. / Raji, I. D. / Gebru, T. (2019): Model Cards for Model Reporting, in: Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT* '19), New York: Association for Computing Machinery, 220–229, DOI: 10.1145/3287560.3287596.

Rawls, J. (1999): A Theory of Justice. Cambridge: Cambridge University Press

Reimann, S. (2024a): Vortrag 'Fairness in KI' vom 14.11.2024, Berlin [unveröffentlichte Quelle].

– (2024b): LinkedIn Korrespondenz vom 20.12.2024 [unveröffentlichte Quelle].

Rudin, C. (2019): Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead, in: Nature Machine Intelligence, Jg. 1 / Nr. 5, 206–215.

Spitznagel, A (2025): Hintergrundgespräch (29.01.2025) [unveröffentliche Quelle].

Uzgalis, W. (2024): John Locke, in: Zalta E. N. / Nodelman, U. (Hrsg.): The Stanford Encyclopedia of Philosophy (Winter 2024 Edition), URL: https://plato.stanford.edu/archives/win2024/entries/locke/ (aufgerufen am: 15/01/2025).

Venkatasubramanian, S. / Alfano, M. (2020): The Philosophical Basis of Algorithmic Recourse, in: Mireille H. / Castillo, C. / Celis, E. / Ruggieri, S. / Taylor, L. / Zanfir-Fortuna, G. (Hrsg.): Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, New York: Association for Computing Machinery, 284–293.

Von Eschenbach, W. J. (2021): Transparency and the Black Box Problem: Why We Do Not Trust AI, in: Philosophy & Technology, Jg. 34 / Nr. 4, 1607–1622.

Wilke, F. (2018): Künstliche Intelligenz diskriminiert (noch), in: Zeit Online, URL: https://www.zeit.de/arbeit/2018-10/bewerbungsroboter-kuenstliche-intelligenz-amazon-frauen-diskriminierung (aufgerufen am: 25/03/2025).

Willaschek, M. / Stolzenberg, J. / Mohr, G. / Bacin, S. (Hrsg.) (2015): Kant-Lexikon, Band 1, Berlin: De Gruyter.