# 1.

# Grundlagen

# Von der Ethik zum Gesetz[1]

## Wie der EU AI Act die Fairness entlang der KI-Wertschöpfungskette formalisiert

Till Klein

### *1. Einführung*

Die Künstliche Intelligenz (KI) durchdringt zunehmend alle Bereiche unseres gesellschaftlichen und wirtschaftlichen Lebens. Von der automatisierten Kreditvergabe in Banken über die Vergabe von Kindergartenplätzen bis hin zur Personalauswahl in Unternehmen – KI-Systeme unterstützen oder treffen Entscheidungen, die das Leben von Menschen maßgeblich beeinflussen. Mit der wachsenden Bedeutung von KI rückt auch die Frage nach der Fairness dieser Systeme immer stärker in den Fokus von Politik, Wirtschaft und Gesellschaft.

Die Europäische Union hat mit der KI-Verordnung (engl. „AI Act") einen wegweisenden regulatorischen Rahmen geschaffen, der erstmals verbindliche Regeln für den Umgang mit Künstlicher Intelligenz etabliert. Der AI Act trat am 1. August 2024 in Kraft und schützt neben der Gesundheit und Sicherheit auch die Grundrechte der Bürgerinnen und Bürger, inklusive dem Schutz vor Diskriminierung. Bei der Umsetzung stellen sich konkrete Fragen:

- Welche Anforderungen sieht der AI Act bezüglich Fairness vor?
- Wie verteilen sich diese Anforderungen auf die KI-Wertschöpfungskette?
- Welche Herausforderungen stellen sich dabei?

---

1 KI wurde für einen initialen Entwurf des Textes auf Grundlage der Folien der Keynote von 2024 verwendet, dann aber verworfen und komplett neu geschrieben. Ein paar wenige Sätze in der Einleitung sind nah am Text der KI.

Dieser Beitrag untersucht, wie sich die neuen rechtlichen Anforderungen im AI Act auf die verschiedenen Stufen der KI-Entwicklung und -Anwendung auswirken – von den Anbietern von Basismodellen über die Entwickler von KI-Systemen bis hin zu den Betreibern und letztendlich den betroffenen Personen.[2] Dabei wird deutlich, dass Fairness in KI-Systemen kein isoliertes technisches Problem darstellt, sondern eine umfassende Aufgabe, die Anstrengungen und Zusammenarbeit der beteiligten Akteurinnen und Akteuren entlang der KI-Wertschöpfungskette erfordert.

## *2. Fairness als Feature*

### *2.1 Wie KI-Systeme diskriminieren*

Fairness in KI-Systemen zu definieren ist komplex und vielschichtig, insbesondere mit Blick auf die interdisziplinäre Natur der Sache. Mathematiker haben ein unterschiedliches Verständnis von Bias im Vergleich zu Soziologen oder Juristen, jedoch sind diese und weitere Rollen häufig bei der Entwicklung und Nutzung von KI-Systemen involviert.

Nach der IEC Norm zu „Bias in KI-Systemen" (ISO/IEC TR 24027:2021) bedeutet Fairness, unparteiisch zu sein und sich ohne Bevorzugung oder Diskriminierung zu verhalten (vgl. ISO/IEC 2021). Der Gegenpol von Fairness ist die Voreingenommenheit (engl. „bias"). Voreingenommenheit beschreibt die systematisch unterschiedliche Behandlung bestimmter Objekte, Personen oder Gruppen im Vergleich zu anderen. Eine ungleiche Behandlung ist per se nicht „schlecht", denn in vielen praktischen Anwendungen werden KI-Systeme explizit entwickelt, um Dinge oder Personen unterschiedlich zu behandeln, etwa in der Qualitätskontrolle, beim Sortieren (z. B. Güteklassen von Obst und Gemüse) oder in der Diagnostik (z. B. bei Krankheitsbildern, die sich in Abhängigkeit von Geschlecht oder Alter unterschiedlich zeigen). Diskriminierung durch KI findet statt, wenn die Voreingenommenheit in einem KI-System zu einer ungerechten oder vorurteilsbehafteten Behandlung von Einzelpersonen, Organisationen, Gruppen oder Gesellschaften führt. Es gibt zahlreiche Beispiele in denen Personen aufgrund ihrer Herkunft, Sprache, Religion oder sexuellen Orientierung benachteiligt behandelt werden, und KI-Systeme bergen das Potential diese Diskriminierung aufzunehmen, zu

[2] Der AI Act spricht von ‚KI-Modellen mit allgemeinem Verwendungszweck' (engl. „General Purpose AI Models", kurz: GPAI-Model).

verstärken und zu skalieren, insbesondere wenn die Trainingsdaten diese Voreingenommenheit widerspiegeln. Allgemein sind die Ursachen von Voreingenommenheit in KI-Systemen ist vielfältig und liegt insbesondere in den frühen Phasen der Systementwicklung (vgl. Norori et al. 2021):

- „Datenbedingter Bias" entsteht durch Trainingsdaten, die nicht repräsentativ für das Phänomen sind, das sie beschreiben sollen, etwa durch überproportional häufiges Vorkommen bestimmter Proben oder Lücken. In diesem Fall lernt das KI-Model diese Verzerrungen und reproduziert sie in seinen Entscheidungen.
- „Algorithmischer Bias" kann durch unausgewogene Klassen oder systematische Fehler im Trainingsprozess entstehen. Auch die Wahl bestimmter Algorithmen, Parameter oder Optimierungsziele kann zu diskriminierenden Ergebnissen führen.
- „Menschlicher Bias" fließt durch gesellschaftliche Vorurteile und Machtungleichgewichte in die Systemgestaltung ein, zum Beispiel durch die Ansichten von Entwicklern, Datenannotatierern[3] und Entscheidungsträgerinnen, die bewusst oder unbewusst ihre eigenen Voreinstellungen mit einbringen.

Das Ergebnis ist ein verzerrter Output, der bestimmte Gruppen systematisch benachteiligt oder bevorzugt. Diese Verzerrungen können sich besonders problematisch auswirken, wenn KI-Systeme in sensiblen Bereichen wie der Kreditvergabe, Personalauswahl oder Strafverfolgung eingesetzt werden.

### *2.2 Ob KI zu mehr Fairness oder Unfairness führt, ist eine Frage der Umsetzung*

Neben der technischen Umsetzung eines KI-Systems hängt es vor allem von der praktischen Umsetzung und dem jeweiligen Kontext ab, ob KI zu mehr oder weniger Fairness führt, was die folgenden Beispiele illustrieren. Ein positives Beispiel ist das KI-System KitaMatch aus Deutschland. Das System adressiert eine gängige Herausforderung junger Eltern bei der Suche nach einem Kita-Platz für den Nachwuchs. Um die eigenen Chancen auf einen Platz zu erhöhen, stellt man Anträge nicht bei einer, sondern bei mehreren Kitas. Die Kitas wiederum beobachten durch die Mehrfachanmeldung eine aufgeblasene Nachfrage, was zu langen Wartelisten, vielen Absagen und

[3] Annotieren bedeutet, Daten mit Anmerkungen oder Notizen zu versehen. Im Bereich KI spricht man umgangssprachlich von „Labeln", also der Markierung von Trainingsdaten durch Menschen, von denen das KI-System lernt.

einem erhöhten Koordinationsaufwand führt. Das KI-System KitaMatch soll hier durch eine Matching-Anwendung Abhilfe schaffen.

> KitaMatch ist ein Verfahren zur fairen, schnellen und transparenten Vergabe von Kitaplätzen, das Eltern und Kitas bei der Vergabe von Betreuungsplätzen optimal miteinander verbindet (KitaMatch 2023).

Die Kitas profitieren unter anderem von „gerichtsfesten" Gründen für die Vergabe, inklusive klare Argumente (gemäß einem Kriterienkatalog), warum andere Kinder priorisiert wurden. Die Jugendämter können sich sicher sein, dass Kinder einheitlich und nachvollziehbar priorisiert wurden. Die Eltern erhalten eine transparente Erklärung, falls das eigene Kind keinen Platz erhält und die standardisierte Vorgehensweise beschleunigt den gesamten Prozess, sodass Eltern sich einen Plan B zurechtlegen können.

Im Kontrast dazu steht das französische KI-System namens „Parcoursup", das zur Vergabe von Universitätsplätzen verwendet wird und dabei zu struktureller Diskriminierung geführt hat (vgl. Federal Anti-Discrimination 2019). Das System gleicht die Bewerbungen von Schulabsolventinnen und Schulabsolventen mit verfügbaren Plätzen an Universitäten auf nationaler Ebene ab, während lokale Universitäten zusätzlich eigene Auswahl-Algorithmen verwenden. Schulabsolventen versenden im Durchschnitt 10–20 Bewerbungen an Universitäten im ganzen Land, wobei sie persönliche Daten angeben müssen, darunter den Wohnort, das Einkommen und die vorherige Schule. Problematisch wurde hier die mangelnde Transparenz: der nationale Algorithmus war öffentlich zugänglich, jedoch nicht lokalen Sortieralgorithmen. Die Angabe von Einkommen und Wohnort führte zu einer systematischen Benachteiligung weniger wohlhabenden Bewerber und Bewerberinnen oder solche aus bestimmten (Vor-)Orten.

Die französische Gleichstellungsbehörde Défenseur des Droits führte eine Untersuchung durch und kritisierte besonders die Verwendung der besuchten Schule als Auswahlkriterium, da dies aufgrund geografischer Lage zu einer Diskriminierung führen kann. Dieser Fall verdeutlicht die Komplexität entlang der KI-Wertschöpfungskette, denn es bedarf ein Zusammenspiel unterschiedlicher Akteurinnen und Akteure und deren Systeme, um die transparente und verantwortungsvolle Nutzung von KI zu gewährleisten. Dieses Beispiel demonstriert außerdem, wie scheinbar neutrale technische Systeme gesellschaftliche Ungleichheiten verstärken können, wenn Transparenz und systematische Fairness-Prüfungen fehlen.

*2.3 Bias in „großen Modellen" multipliziert sich entlang der KI-Wertschöpfungskette*

Die Popularität „großer Modelle", die meist von großen Tech-Firmen wie OpenAI, META, Amazon oder Google zur Verfügung gestellt werden, stellen eine besondere Herausforderung für Transparenz und Fairness dar, weil sie für den Mangel an Transparenz kritisiert werden und gleichzeitig bei zigtausenden nachgelagerten Anbietern (engl. „Downstream-Provider") zum Einsatz kommen. Ein plakatives Beispiel für Bias in Basismodellen ist die Studie „Which Humans?" von Atari et al. aus dem Jahr 2023. Die Untersuchung beginnt mit der Prämisse, dass die Anbieterinnen und Anbieter dieser Modelle gerne von der Demokratisierung von KI sprechen und erklären, dass die neuen KI-basierten Fähigkeiten für „alle Menschen" sind. Die Autoren des Papers fragen „which humans?" („welche Menschen") denn die Ergebnisse zeigen, dass die getesteten Basismodellen einen kulturellen Bias innehaben und die Werte und Konzepte von Menschen in Abhängigkeit ihrer Herkunft mehr oder weniger gut widerspiegeln. Konkret: je „näher" die Kultur eines Landes an der Kultur der USA liegt, desto besser resonieren die Ergebnisse der Basismodelle mit den Nutzenden. Zum Beispiel sehen sich Nutzenden aus Kanada und Australien sehr gut in den Ergebnissen, während das Gegenteil für Personen aus Ägypten, Jordanien oder Pakistan gilt (vgl. Abbildung 1).

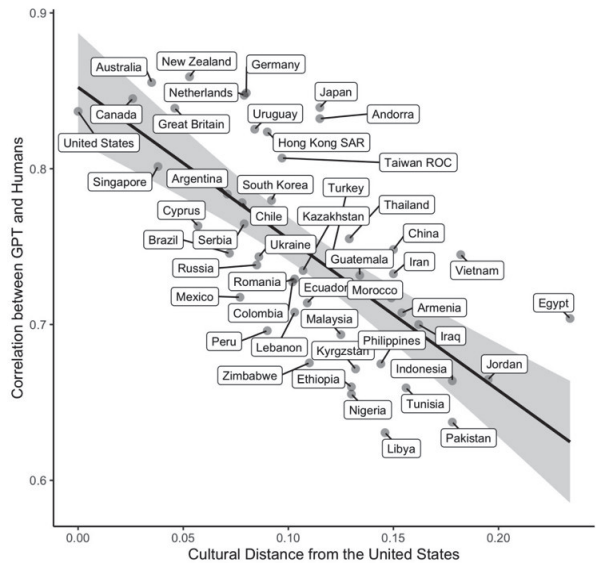


ABBILDUNG 1: KULTURELLER BIAS (QUELLE: ATARI ET AL. 2023: 11)

Sofern Bias in einem KI-Modell bekannt und transparent kommuniziert ist, können nachgelagerte Akteurinnen und Akteure nach einem geeigneten Umgang suchen und zum Beispiel ein anderes Modell verwenden oder Mitigationsmaßnahmen umsetzen. Hier liegt eine Herausforderung, denn erstens ist es häufig unklar ob bzw. welche Schwächen an Basismodellen bekannt sind (selbst bei den eigenen Anbietern, siehe „Emerging Abilities" [vgl. Berti et al. 2025]) und zweitens werden die Model-Anbieter häufig für einen Mangel an Transparenz kritisiert. Der AI Safety Report bewertet führende Basismodelle anhand bestimmter Kriterien und kommt zu folgendem Fazit (übersetzt aus dem Englischen): Trotz des wachsenden internationalen Konsenses über die Risiken der KI und der zunehmenden Belege für rasante Fortschritte bei den Fähigkeiten warnen Experten, dass die Kluft zwischen technologischen Ambitionen und Sicherheitsvorkehrungen immer größer wird. Unternehmen streben nach künstlicher allgemeiner Intelligenz und sagen voraus, dass sie innerhalb dieses Jahrzehnts übermenschliche Leistungen erzielen werden. Doch wie ein Prüfer feststellte, „hat keines der Unternehmen einen kohärenten, umsetzbaren Plan" zur Kontrolle solcher Systeme (vgl. Future of Life Institute 2025: 21).

Ungeachtet der oben genannten Mängel hat die Nutzung von generativer KI seit ChatGPT rasant zugenommen, weil wesentliche Adaptions-barrieren gesunken sind. Die Benutzerschnittstelle ist einfach zu bedienen, Ein- und Ausgabe erfolgen in natürlicher Sprache, und Rechenleistung und Speicher werden über die Plattform im Hintergrund abgedeckt. Laut einer Studie des Start-Up Verbands (vgl. Startup Verband 2024) haben 43% der befragten Unternehmen GenAI regelmäßig im Einsatz und weitere 35% nutzen es im operativen Alltag. Mit jedem Einsatz, jeder Adaption, werden die Verzerrungen und Mängel der generativen Modelle weiter multipliziert und verbreitet.

### *2.4 Fairness in KI als Weg zu mehr Wettbewerbsfähigkeit?*

„Ohne Vertrauen, keine Nutzung" (engl. „no trust, no use") – so lautet eine gängige Formel der Befürworterinnen und Befürworter von vertrauenswürdiger KI. Wenn wir einem KI-System nicht trauen können, zum Beispiel für die Korrektheit oder Nachvollziehbarkeit der Ergebnisse, dann werden wir es kaum (für kritische Anwendungen) einsetzen. Der Umkehrschluss suggeriert, dass Vertrauenswürdigkeit ein wesentlicher Faktor ist, um die Adaption von KI zu erhöhen, also: „mehr Vertrauen, mehr Nutzung".

Eine Umfrage des Digital-Verbands Bitkom (2024) zeigt, dass ein überschaubarer Anteil von 9% der befragten 602 Unternehmen GenAI bereits nutzen (vgl. Bitkom 2024). Weitere 18% planen den Einsatz und weitere 19% können es sich vorstellen. Demgegenüber stehen die privaten Nutzer und Nutzerinnen, bei denen bereits 30% der deutschen Gesamtbevölkerung allein ChatGPT aktiv nutzen, wie eine Umfrage des Nürnberg Institut für Marktentscheidungen e.V. berichtet (vgl. Kaiser et al. 2024). Insbesondere bei etablierten Unternehmen gibt es Zurückhaltung bei Investitionen in die Nutzung von GPAI-Modellen, unter anderem wegen Bedenken bei ihrer Zuverlässigkeit. Diese Daten deuten auf einen Mangel an Vertrauen gegenüber generativer KI in etablierten Unternehmen.

Zahlreiche Umfragen setzen hier an und beleuchten die Hindernisse für die Nutzung von KI bzw. generativer KI, wobei die Leistungsfähigkeit der Modelle ein wiederkehrender Aspekt ist. Unter den Top 5 Gründen der Bitkom-Umfrage nennen 65% der Unternehmen „Schlechte Qualität der Ergebnisse" als Hemmnis für den Einsatz von generativer KI (vgl. Bitkom 2024). Eine Umfrage der Expertenkommission Forschung und Innovation zeigt, dass „Bedenken hinsichtlich Reife und Zuverlässigkeit von KI" ein wesentliches Hindernis für den Einsatz von KI in Unternehmen ist (vgl. EFI 2024). Auch bei der Nutzung im privaten Umfeld unterstreicht das Eurobarometer zur Digitalen Dekade der EU, dass sich die Mehrheit der Befragten digitale Dienstleistungen wünschen, die besser auf die persönlichen Bedürfnisse abgestimmt sind.

In dem Korridor zwischen „Bias in GPAI Modellen" und „Keine Nutzung von GenAI wegen schlechter Qualität der Ergebnisse" liegt die Chance robuste Modelle zu bauen, die alle Nutzergruppen „gleich gut" behandeln, unabhängig von ihrem Herkunftsland, die zuverlässige Ergebnisse erzeugen und die den Mensch in den Mittelpunkt stellen. Wenn dies gelingt, können Fairness, und andere Aspekte vertrauenswürdiger KI, ein Erfolgsfaktor für die breitere Nutzung von KI in Unternehmen und im persönlichen Umfeld sein. Die Beispiele von KitaMatch und Parcoursup aus Frankreich zeigen, dass dafür alle Akteurinnen und Akteure in der KI-Wertschöpfungskette eine Rolle spielen, vom GPAI-Anbieter, über die nachgelagerten Anbieter, bis hin zum Endnutzer. Der AI Act hat diese unterschiedlichen Rollen aufgegriffen und sieht konkrete Pflichten und Rechte für sie vor.

## *3. Von der Ethik zum Gesetz: Fairness in der europäischen KI-Verordnung*

### *3.1 Ethische Richtlinien als Säule des AI Acts*

Der AI Act wird mitunter als formalisierte Ethik bezeichnet, weil die neuen Regularien auf ethischen Prinzipien aufbauen und konkrete Anforderungen für die ethische Entwicklung und Nutzung von KI beinhalten. Diese Entwicklung ist konsistent mit dem Vorgehen der EU, denn bereits 2019 hat eine von der EU beauftragte Expertengruppe den Begriff „vertrauenswürdige KI" als Leitbild formuliert (vgl. EU Kommission 2019). Vertrauenswürdige KI hat drei wesentliche Merkmale (übersetzt aus dem Englischen):

1. Rechtmäßig – Einhaltung aller geltenden Gesetze und Vorschriften
2. Ethisch – Einhaltung ethischer Grundsätze und Werte
3. Robust – aus technischer Sicht und unter Berücksichtigung des sozialen Umfelds

### *3.2 Schutz von Grundrechten als explizites Ziel*

Beim Blick in den AI Act finden sich klare Hinweise für eine ethische und Mensch-zentrierte Ausgestaltung des Regelwerkes, beginnen bei Artikel 1, dem Gegenstand:

> (1) Zweck dieser Verordnung ist es, [...] die Einführung einer auf den Menschen ausgerichteten und vertrauenswürdigen künstlichen Intelligenz (KI) zu fördern und gleichzeitig ein hohes Schutzniveau in Bezug auf Gesundheit, Sicherheit und die in der Charta verankerten Grundrechte, [...] vor schädlichen Auswirkungen von KI-Systemen in der Union zu gewährleisten und die Innovation zu unterstützen.

Die KI-Verordnung definiert damit den Schutz der Grundrechte als eine zentrale Säule und verfolgt dabei einen menschenzentrierten Ansatz für vertrauenswürdige KI. Zusätzlich zielt der Gesetzgeber darauf ab, ein hohes Schutzniveau für andere Aspekte wie Gesundheit, Sicherheit, Demokratie und Rechtsstaatlichkeit zu gewährleisten. Neben den Schutzzielen soll mit dem AI Act Innovation und wirtschaftliche Entwicklung gefördert werden, d.h. es geht nicht um Regulierung ODER Innovation, sondern um die balancierte Symbiose der vermeintlich im Konflikt stehenden

Ziele, wobei die die Grundwerte der Union, im Sinne der Grundrechte der Europäischen Union („Charta"), als normative Leitplanken fungieren. Die Charta ist Teil der Europäischen Verträge und gibt den Bürgerinnen und Bürgern der EU-Rechte, die unter anderem auf Fairness und Gleichbehandlung abzielen (vgl. Abbildung 2) (vgl. Europäische Union 2000):

| Artikel Nr. | Titel |
|---|---|
| 21 | Nichtdiskriminierung |
| 23 | Gleichheit von Männern und Frauen |
| 24 | Rechte des Kindes |
| 25 | Rechte älterer Menschen |
| 26 | Integration von Menschen mit Behinderung |

ABBILDUNG 2: BEISPIELHAFTE AUSWAHL VON GRUNDRECHTEN DER EU ZUM THEMA FAIRNESS UND GLEICHBEHANDLUNG (QUELLE: EUROPÄISCHEN UNION 2000: 13 F.)

### *3.3 Ethische Richtlinien als Säule des AI Acts*

Die erfolgreiche Überführung ethischer Prinzipien zum Gesetz attestieren auch wissenschaftliche Aufsätze, etwa von Nathalie Smuha von der KU Leuven in Belgien (vgl. Smuha 2024). Ihre Analyse zeigt auf, dass die Empfehlungen der sogenannten High-Level Expert Group (vgl. EU Kommission 2019) zu vertrauenswürdiger KI ihren Weg in den Text des AI Acts gefunden haben. Dazu zählen vier zentrale Grundsätze, inklusive Fairness (die anderen drei Grundsätze sind: Achtung der menschlichen Autonomie, Schadensverhütung und Erklärbarkeit). Konkret schlagen sich die ethischen Richtlinien wieder in den Anforderungen an Hochrisiko-KI-Systeme (Artikel 8–15, z. B. im Bereich Data Governance) und bei der Auswahl verbotener KI-Praktiken (Artikel 5). In der Summe dienten die ethischen Richtlinien als normativer Kompass für den AI Act und stellten die Grundlage für den freiwilligen Verhaltenskodex.

Diese nachfolgende Sektion behandelt die Frage, welche Anforderungen an Fairness entlang der KI-Wertschöpfungskette im AI Act zu finden sind. Für diesen Zweck betrachten wir ein fiktives Szenario in der Finanzbranche und untersuchen anhand dessen, welche möglichen Änderungen zu erwarten sind.

## 4. *Fairness entlang der KI-Wertschöpfungskette*

### *4.1 Ein exemplarisches Szenario: Kreditwürdigkeitsprüfung (Credit Scoring)*

Kreditanträge können in den unterschiedlichsten Lebenslagen notwendig sein: das erste Auto, eine Immobilie, eine Unternehmensgründung oder der große Wunsch, der noch offen war. Ein wesentlicher Schritt auf der Seite des Kreditinstitutes ist die Bonitätsprüfung, also die Prüfung, ob und zu welchen Bedingungen ein Kredit vergeben werden kann. Im Hintergrund steht die Frage der Ausfallwahrscheinlichkeit: Wird der Kreditnehmer den Betrag vollständig und fristgerecht zurückzahlen können? Um diese Prüfung zu automatisieren und zu beschleunigen, kommen immer häufiger KI-Systeme zum Einsatz, die anhand bestimmter Eingabewerte, die Ausfallwahrscheinlichkeit des Antragstellers bewerten und damit die Entscheidung zur Kreditvergabe wesentlich beeinflussen können. Stellen wir uns vor, ein junges Paar stellt einen Kreditantrag (z. B. für den Kauf einer Immobilie) und die Finanzberaterin erklärt, dass die Kreditwürdigkeit mithilfe eines KI-Systems ermittelt wurde. Bei der Besprechung der Konditionen des Kredits, die überraschend ungünstig ausfallen, beschleicht das Paar ein seltsames Gefühl und es fragt nach, wie der Bonitätswert zustande kam.

Mit Blick auf den AI Act, ist das Kreditinstitut in diesem Beispiel der Betreiber des KI-Systems und das Paar sind die betroffenen Personen. Der Anbieter des KI-Systems ist möglicherweise eine andere Entität, z. B. ein FinTech Startup, das unter anderem ein großes Sprachmodell[4] nutzt, mit den Informationen aus dem Internet und den Sozialen Medien über die Antragsteller ausgewertet werden, etwa mit Blick auf „gefährliche Hobbies" oder andere Verhaltensweisen, die die Kreditwürdigkeit beeinflussen könnten. Basierend auf der KI-Wertschöpfungskette, bestehend aus GPAI-Model Anbieter, KI-System Anbieter, KI-System Betreiber und betroffener Personen, stellt sich die Frage: Welche Pflichten oder Rechte des AI Acts tragen zu mehr Fairness für die betroffenen Personen bei?

Die folgenden Absätze betrachten verschiedene Rollen entlang der KI-Wertschöpfungskette und (a) zeigen zentrale Anforderungen aus dem AI Act zu Fairness auf und (b) weisen auf offene Punkte hin (Stand Mitte 2025).

---

[4] Nehmen wir an, beim Sprachmodell handelt es sich um ein „KI-Modell mit allgemeinem Verwendungszweck" (GPAI-Model) im Sinne des AI Acts.

Die betroffene Person ist die Person, die von dem Ergebnis oder einer Entscheidung eines KI-Systems betroffen ist. Im oben genannten Beispiel ist es das Paar, denn die Bedingungen für ihren individuellen Kredit werden von dem KI-System für die Kreditwürdigkeitsprüfung beeinflusst. Der AI Act enthält Rechtsbehelfe (Kapitel 9, Abschnitt 4), die die Position von betroffenen Personen stärken. Konkret geht es um das Recht auf Erläuterung (Artikel 86), wonach betroffene Personen von Hochrisiko-KI-Systemen im Fall eines begründeten Verdachts das Recht haben „vom Betreiber eine klare und aussagekräftige Erläuterung zur Rolle des KI-Systems im Entscheidungsprozess und zu den wichtigsten Elementen der getroffenen Entscheidung zu erhalten". Falls Grund zur Annahme besteht, dass gegen Bestimmungen des AI Acts verstoßen wurde, können gemäß Artikel 85 natürliche und juristische Personen „bei der betreffenden Marktüberwachungsbehörde Beschwerden einreichen". Im Sinne des Beispiels könnte das Paar vom Kreditinstitut eine Erläuterung verlangen, da das KI-System zur Kreditwürdigkeitsprüfung gemäß Artikel 6 in Kombination mit Anhang III wahrscheinlich ein Hochrisiko-KI-System ist.

Die durch den AI Act eingeführten Rechte für betroffene Personen verringern das Risiko, mit „unfairen" KI-Systemen konfrontiert zu werden, weil Anbieter mit entsprechenden Rückfragen rechnen müssen. Gleichzeitig können sich betroffene Personen auf ihr Recht berufen, bzw. im Schadensfall eine Wiedergutmachung verlangen. Inwiefern Fairness tatsächlich gestärkt wird, hängt jedoch von der Umsetzung bzw. Durchsetzung dieser Rechte ab, da es in der Praxis noch offene Fragen gibt. Hier eine Auswahl:

- Begriffsdefinition: Der Begriff „betroffene Person" ist nicht im AI Act definiert, wodurch ein Interpretationsspielraum entsteht, der im Einzelfall unterschiedlich ausgelegt werden kann. Zum Beispiel ist unklar, welches Tatbestandsmerkmal erfüllt sein muss, damit eine Person als „betroffen" gilt.
- Bewusstsein: Nur wer die eigenen Rechte kennt, kann sie geltend machen. Jedoch sind die neuen Rechte durch den AI Act weitgehend unbekannt unter den potenziell betroffenen Personen (insbesondere Verbraucherinnen und Verbraucher). Bei der Datenschutz Grundverordnung (DSGVO) hat es Jahre gedauert, ein gewisses Bewusstsein in der Gesellschaft aufzubauen (vgl. Rughinis et al. 2019).
- Durchsetzung: Jeder Mitgliedstaat muss gemäß Artikel 70 bis zum 02. August 2025 eine Marktaufsichtsbehörde benennen und öffentlich bekannt geben, wie diese Behörde über

elektronische Wege erreicht werden kann. Jedoch ist diese Behörde Mitte August 2025 für Deutschland noch nicht offiziell benannt worden, das heißt, betroffene Personen können ihr „Recht auf Beschwerde“ noch nicht ausüben.

### *4.3 Fairness und Betreiber von KI-Systemen*

Betreiber von KI-Systemen sind natürliche oder juristische Personen, die ein KI-System in eigener Verantwortung verwenden, siehe Artikel 3 (4) AI Act. Im Beispiel der Kreditwürdigkeitsprüfung ist das Kreditinstitut der Betreiber, weil die Finanzberaterin das KI-System in ihrer Rolle als Angestellte und im Auftrag des Arbeitgebers einsetzt.

Der AI Act enthält mehrere Anforderungen an Betreiber, die zu mehr Fairness für die betroffenen Personen führen sollten. Erstens sind Betreiber von bestimmten Hochrisiko-KI-Systemen, darunter auch solche für Kreditwürdigkeitsprüfungen, angehalten, eine Grundrechtefolgenabschätzung vor der Inbetriebnahme durchzuführen, siehe Artikel 27. Die zentrale Frage ist, welche Auswirkungen die Verwendung des KI-Systems auf die Grundrechte haben kann, insbesondere mit Blick auf Schadensrisiken für betroffene Personen. Zweitens definiert Artikel 26 spezifische Pflichten für Betreiber von Hochrisiko-KI-Systemen, zum Beispiel die Nutzung des KI-Systems gemäß der Betriebsanleitung des Anbieters, die Benennung einer menschlichen Aufsicht über das KI-System, sowie die Pflicht, Vorfälle mit dem KI-System an den Anbieter und die zuständigen Behörden zu melden. Außerdem müssen Arbeitnehmervertreter vor der Inbetriebnahme des KI-Systems darüber informiert werden.

Bei der Einhaltung dieser Pflichten hat das Paar im Beispiel der Kreditwürdigkeitsprüfung eine höhere Chance auf eine faire Behandlung, weil mögliche Grundrechtsverletzungen vor der Inbetriebnahme ermittelt und mitigiert wurden. Auch die Finanzberaterin sollte mit dem KI-System vertraut sein und unregelmäßige Ergebnisse erkennen bzw. korrigieren können. Dennoch gibt es auch hier offene Punkte in der praktischen Umsetzung:

- Methodik: Zur effektiven Durchführung einer Grundrechtefolgenabschätzung von KI-Systemen bedarf es allgemein anerkannter Methoden, jedoch gibt weder der AI Act diese Methode vor, noch enthält er einen Auftrag an das AI Office oder anderen Akteurinnen und Akteure, diese Methode zu entwickeln.[5]

[5] Das dänische Institut für Menschenrechte bietet eine umfassende Methode für Grundrechtefolgenabschätzungen im Digitalen Kontext (vgl. Danish Institute for Human Rights 2025).

- Vorfälle: Im Gegensatz zu Vorfällen bezüglich Gesundheit und Sicherheit sind Vorfälle an den Grundrechten schwieriger zu erkennen, da diese mitunter subtil, implizit oder akkumulativ auftreten, zum Beispiel eine täglich wiederkehrende Diskriminierung durch Kaufempfehlungssystem. Vor allem Betreiber brauchen hier ein „geschultes Auge", um Vorfälle in Einklang mit dem AI Act zu erkennen, zu bewerten und gegebenenfalls zu melden.

### *4.4 Fairness und Anbieter von KI-Systemen*

Anbieter von KI-Systemen sind natürliche und juristische Personen, die KI-Systeme entwickeln oder entwickeln lassen, den Verwendungszweck bestimmen und/oder ein KI-System unter dem eigenen Namen verfügbar machen, siehe Artikel 3 (1) und Artikel 25 für Details. Ein Großteil der Pflichten im AI Act richtet sich an Anbieter, weil deren Entscheidungen während der Entwicklung ganz maßgeblich die Leistung und Sicherheit eines KI-Systems in der Nutzung beeinflussen. Im Beispiel der Kreditwürdigkeitsprüfung ist das Fintech-Startup der Anbieter.

Die Pflichten für Anbieter sind abhängig von der Risikoeinstufung des jeweiligen KI-Systems. Hochrisiko-KI-Systeme sind mit umfangreichen Anforderungen belegt, weil diese ein Konformitätsbewertungsverfahren durchlaufen müssen, bevor sie in Verkehr gebracht oder in Betrieb genommen werden dürfen (Artikel 43). Im Mittelpunkt stehen dabei die „Anforderungen an Hochrisiko-KI-Systeme" (Kapitel III, Abschnitt II, Artikel 8-15), wozu auch Risikomanagement, Data Governance, menschliche Aufsicht Genauigkeit, Robustheit und Cybersicherheit zählen. KI-Systeme mit begrenztem Risiko unterliegen sogenannten Transparenzpflichten (Artikel 50), die sicherstellen sollen, dass KI-basierte Inhalte oder interaktive Systeme (z. B. Chatbots) als solche erkennbar sind. Niedrigrisiko-KI-Systeme unterliegen keinen besonderen Pflichten, aber Anbieter werden ermutigt, einen freiwilligen Verhaltenskodex zu befolgen (Artikel 95).

In diesen Pflichten liegt ein zentraler Hebel, um mehr Fairness für betroffene Personen sicherzustellen. Die sorgfältige Auswahl von Trainingsdaten, Mechanismen für Transparenz und Reproduzierbarkeit sowie umfassende Informationen an die späteren Nutzer und Nutzerinnen des KI-Systems unterstützen den kontrollierten und verantwortungsvollen Einsatz von KI. Wie auch bei den anderen Stufen in der KI-Wertschöpfungskette gibt es auf der Ebene von Anbietern offene Punkte:

- Standards: Die Verwendung von Standards ist grundsätzlich freiwillig, aber laut Artikel 40 im AI Act bilden die sogenannten harmonisierten Standards die Grundlage für die Konformitätsvermutung, also die Konformität mit den „Anforderungen an Hochrisiko-KI-Systeme" (Artikel 8-15). Nach aktuellen Berichten (Stand August 2025) wird sich die Veröffentlichung der harmonisierten Standards deutlich verspäten, was auch deren praktische Umsetzung inklusive positiver Effekte für mehr Fairness in die Zukunft verschiebt.
- Marktüberwachung: Kompetente Behörden sind die Voraussetzung für die effektive Durchsetzung dieser Pflichten, jedoch ist mit Blick auf Umsetzungsgeschwindigkeit des AI Acts in den Mitgliedstaaten und die allgemein angespannte Haushaltslage unklar, wann und in welchem Umfang die jeweiligen Behörden bzw. deren Angestellte in angemessenen Umfang dafür befähigt und beauftragt sind.

### *4.5 Fairness und Anbieter von KI-Modellen mit allgemeinem Verwendungszweck*

Anbieter von GPAI-Modellen sind natürliche und juristische Personen, ein KI-Modell mit allgemeinem Verwendungszweck (engl. „General Purpose AI Model") entwickeln oder entwickeln lassen oder ein GPAI-Modell unter dem eigenen Namen verfügbar machen, siehe Artikel 3 (1) und die Richtlinien des KI-Büros der EU Kommission für Anbieter von KI-Modell mit allgemeinem Verwendungszweck (vgl. AI Office 2025a). Die meisten bekannten Anbieter haben ihren Sitz außerhalb der EU, zum Beispiel Google DeepMind, Meta, Amazon, X oder Microsoft. Im Beispiel der Kreditwürdigkeitsprüfung könnte das Fintech Startup ein GPAI-Model als eine Komponente des spezifischen KI-Systems integrieren.

Die Pflichten im AI Act für Anbieter von GPAI-Modellen orientieren sich an der Leistungsfähigkeit des Modells und in Abhängigkeit, ob das Model „open source" oder proprietär ist. Die geringsten Pflichten gelten nach Artikel 53 (2) für „freie und offene" GPAI-Modelle, die eine Zusammenfassung ihrer Trainingsdaten veröffentlichen und europäisches Recht für Urheberrechte beachten müssen. Die nächste Stufe sind proprietäre GPAI-Modelle, die zusätzlich technische Dokumentation über Test- und Trainingsprozess veröffentlichen und relevante Informationen für nachgelagerte Anbieter (wie das Beispielhafte FinTech Startup) zur Verfügung stellen müssen. Die meisten Pflichten gelten für GPAI-Modelle mit systemischem Risiko, also Modelle, die über überdurchschnittliche Kapazitäten oder eine besonders hohe Reichweite verfügen. An-

bieter solcher Modelle müssen laut Artikel 55 zusätzlich einen Risikomanagementprozess umsetzen, Maßnahmen für Cybersecurity befolgen und einen Prozess für die Bearbeitung von Vorfällen betreiben.

All diese Pflichten zielen darauf ab, die Qualität und Robustheit von GPAI-Modellen zu verbessern, was einen positiven Effekt auf den letztendlichen Einsatz mit betroffenen Personen haben sollte, weil die Nachweise Sicherheit und Transparenz für nachgelagerte Anbieter schaffen. Am 02. August hat das KI-Büro den freiwilligen Verhaltenskodex für GPAI-Modelle (vgl. ebd. 2025b) veröffentlicht, in dem konkrete Umsetzungshinweise für Artikel 53 und 55 zu finden sind. Bevor diese Anforderungen einen Effekt zeigen, gilt es:

- Einhaltung: Zwar haben sich (Stand August 2025) mehrere große GPAI-Anbieter entschlossen, dem Kodex per Selbstverpflichtung zu folgen, dennoch steht aus, wie der Kodex konkret in der Praxis umgesetzt wird. Eine wesentliche Herausforderung dabei sind die unterschiedlichen Informationsbedarfe der zigtausend KI-Anwendungen bei nachgelagerten Anbietern, die auf die Auskünfte der GPAI-Anbieter angewiesen sind.
- Durchsetzung: Die Durchsetzung der Regeln für GPAI-Modelle wird teilweise vom KI-Büro übernommen (Artikel 75) und birgt zwei Herausforderungen. Erstens ist der Kodex freiwillig, sodass sich das KI-Büro im Zweifel auf Artikel 53 und 55 berufen kann, aber nicht auf den Kodex an sich. Zweitens hat die Durchsetzung eine wirtschaftliche und geopolitische Dimension, aufgrund der großen Marktanteile der US-amerikanischen GPAI-Anbieter.

## *5. Zwischenfazit*

An dieser Stelle eignet sich ein Zwischenfazit, weil der Umsetzungsprozess des AI Acts in vollem Gang ist und es regelmäßig neue Fortschritte gibt. Mit Blick auf die eingangs gestellten Fragen zeigt dieser Beitrag, dass der AI Act konkrete Anforderungen an Fairness von KI-Systemen enthält und dass diese Anforderungen auf ethischen Prinzipien beruhen. Diese Beobachtung unterstützt die Ansicht, den AI Act als formalisierte Ethik zu bezeichnen. Die Anforderungen an Fairness verteilen sich über die verschiedenen Stufen der KI-Wertschöpfungskette: Betroffene Personen erhalten durch den AI Act neue Rechte (Recht auf Erläuterung, Recht auf Beschwerde), hingegen sehen sich Betreiber und Anbieter von KI-Systemen und GPAI-Modellen mit neuen Pflichten konfrontiert. Der Umfang der Pflichten folgt dabei stets dem Risikobasierten Ansatz: bei KI-Systemen ist die

Risikoeinstufung maßgeblich (Hochrisiko, begrenztes Risiko, niedriges Risiko) und bei GPAI-Modellen hängt es von der Klassifizierung mit oder ohne systemische Risiken ab.

Zwar stellen die neuen Rechte und Pflichten eine Tendenz zu „mehr Fairness in KI" in Aussicht, aber die tatsächliche Wirkung des AI Acts hängt von der Um- und Durchsetzung der neuen Regeln ab, wobei sich einige Herausforderungen stellen. Betroffene Personen können nur effektiv vor Diskriminierung geschützt werden bzw. sich davor schützen, wenn klar ist, wer genau damit gemeint ist und wenn diese Personen Kenntnis über ihre neuen Rechte im AI Act haben. Betreiber von KI-Systemen müssen interne Methoden und Fähigkeiten aufbauen. Einerseits, um potenzielle Diskriminierung im Rahmen von einer Grundrechtefolgenabschätzungen zu identifizieren und mitigieren, andererseits um Vorfälle während der Nutzung als solche zu erkennen und zu behandeln. Anbieter von KI-Systemen haben den „Löwenanteil" der Pflichten auf ihrer Seite und befinden sich gleichzeitig in einem Spannungsfeld. Die Pläne, sich durch KI einen Wettbewerbsvorteil zu verschaffen, werden durch Verzögerungen bei der Veröffentlichung der harmonisierten Standards und der Etablierung der nationalen Marktüberwachung gedämpft. Der Verhaltenskodex für GPAI-Modelle (Seit Juli 2025) ist eine wichtige Referenz für mehr Transparenz und klare Verantwortlichkeiten bei Basismodellen, aber seine freiwillige Natur birgt Herausforderungen im Bereich der Einhaltung von Anbietern und Durchsetzung von Behörden.

## *6. Ausblick*

Es bleibt weiter spannend in der Arena um KI-Fairness, denn das Feld ist sehr dynamisch und viele Akteurinnen und Akteure, mit teilweise unterschiedlichen Vorstellungen, sind involviert. Die Umsetzung des AI Acts schreitet gemäß der Fristen in Artikel 113 voran und der 2. August 2025 war der jüngste Stichtag. Seither gelten zum Beispiel die Regelungen für die Nationale Aufsicht (Artikel 28), die Pflichten für Anbieter von GPAI-Modellen (Artikel 53 und 55), die Governance auf EU-Ebene (Artikel 64, 65) und die Sanktionen (Artikel 99). Der nächste große Meilenstein ist der 02. August 2026, wenn das Gros der Pflichten in die Anwendung kommt, insbesondere die Pflichten für bestimmte Hochrisiko-KI-Systeme sowie Transparenzanforderungen für KI-Systeme mit begrenztem Risiko.

Auf EU-Ebene gibt es verschiedene Handlungsstränge, die einen Einfluss auf die Fairness für KI-Systeme bzw. digitale Dienste im Allgemeinen haben können. Bereits 2024 hat die EU

Kommission einen Entwurf für den „Digital Fairness Act“ (vgl. EU Kommission 2024) präsentiert und sammelt bis zum 24. Oktober 2025 Feedback und Nachweise („evidence“) für die mögliche Wirkung, die ein solcher Act haben könnte. Das Ziel ist es, Verbraucherinnen und Verbraucher in der Digitalwirtschaft besser zu schützen. Parallel dazu arbeitet die EU Kommission an einem sogenannten „Omnibus Package“ (vgl. EU Kommission 2025) mit dem Ziel, bestehende Regulierung im digitalen Sektor zu vereinfachen („Simplification“), wobei es insbesondere um die Verschlankung und Integration geht, um langfristig die Wettbewerbsfähigkeit von Unternehmen zu stärken.

Fairness im Bereich KI kann zu einem „win-win“ werden, wenn die gängige Rhetorik a la „Innovation versus Regulierung“ abgelöst und durch „mehr Fairness führt zu mehr Innovation“ ersetzt wird. Umfrageergebnisse zeigen, dass sowohl Unternehmen als auch Privatpersonen KI unter anderem auch deshalb nicht nutzen, weil es Bedenken bezüglich der Zuverlässigkeit und Vertrauenswürdigkeit gibt. Sollte es auch mit Hilfe des AI Acts gelingen, die Qualität von KI-Systemen zu erhöhen, kann das die Adaption von KI vorantreiben. Als Resultat hieße das mehr Schutz für betroffene Personen und mehr Wettbewerbsfähigkeit und Produktivität in Unternehmen.

Bis dahin ist viel zu tun. Gemäß dem Sprichwort “it takes a village to raise a child” zeigt dieser Beitrag “it takes an AI value chain to get a fair AI system”. Es braucht Anstrengungen, Ideen und Zusammenarbeit, um eine faire digitale Zukunft in der EU zu erreichen. Der AI Act geht den Schritt von der Ethik zum Gesetz, aber es ist eher ein Marathon als ein Sprint, bis das Gesetz zur Realität wird.

## *Literaturverzeichnis*


AI Office (2025a): Guidelines on the Scope of Obligations for Providers of General-Purpose AI Models Under the AI Act, URL: https://digital-strategy.ec.europa.eu/en/library/guidelines-scope-obligations-providers-general-purpose-ai-models-under-ai-act (aufgerufen am: 05/08/2025).

– (2025b): The General-Purpose AI Code of Practice, URL: https://digital-strategy.ec.europa.eu/en/policies/contents-code-gpai (aufgerufen am: 05/08/2025).

Atari, M. / Xue, M. J. / Park, P. S. / Blasi, D. E. / Henrich, J. (2023): Which Humans?, Cambridge: Havard University Press, URL: https://scholar.harvard.edu/sites/scholar.harvard.edu/files/henrich/files/which_humans_09222023.pdf (aufgerufen am: 05/08/2025).

Berti, L / Giorgi, F. / Kasneci, G. (2025): Emergent Abilities in Large Language Models: A Survey, URL: https://arxiv.org/abs/2503.05788 (aufgerufen am: 05/08/2025).

Bitkom (2024): Künstliche Intelligenz in Deutschland, URL: https://www.bitkom.org/sites/main/files/2024-10/241016-bitkom-charts-ki.pdf (aufgerufen am: 05/08/2025).

Danish Institute for Human Rights (2025): Human Rights Impact Assessment of Digital Activities, URL: https://www.humanrights.dk/publications/human-rights-impact-assessment-digital-activities (aufgerufen am: 06/08/2025).

EFI (2024): Gutachten zu Forschung, Innovation und technologischer Leistungsfähigkeit Deutschlands 2024, Berlin: EFI.

EU Kommission (2019): Ethics Guidelines for Trustworthy AI, URL: https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai (aufgerufen am: 05/08/2025).

– (2024): Digital Fairness Act, URL: https://ec.europa.eu/info/law/better-regulation/have-your-say/initiatives/14622-Digital-Fairness-Act_en (aufgerufen am: 05/08/2025).

– (2025): Commission Proposes to Cut Red Tape and Simplify Business Environment, URL: https://commission.europa.eu/news-and-media/news/commission-proposes-cut-red-tape-and-simplify-business-environment-2025-02-26_en (aufgerufen am: 07/08/2025).

Europäische Union (2000): Charta der Grundrechte der Europäischen Union, URL: https://www.europarl.europa.eu/charter/pdf/text_de.pdf (aufgerufen am: 05/08/2025).

Federal Anti-Discrimination Agency (2019): Risks of Discrimination through the Use of Algorithms, URL: https://www.antidiskriminierungsstelle.de/EN/homepage/_documents/download_diskr_risiken_verwendung_von_algorithmen.pdf?__blob=publicationFile&v=1 (aufgerufen am: 05/08/2025).

Future of Life Institute (2025): AI Safety Report 2025, URL: https://futureoflife.org/wp-content/uploads/2025/07/FLI-AI-Safety-Index-Report-Summer-2025.pdf (aufgerufen am: 06/08/2025).

ISO / IEC (2021): ISO/IEC TR 24027:2021 Information Technology – Artificial intelligence (AI) – Bias in AI Systems and AI Aided Decision Making.

Kaiser, C. / Buder, F. / Biró, T. (2024): ChatGPT und Co. im Alltag: Nutzung, Bewertung und Zukunftsvisionen. Ein Drei-Länder-Vergleich, NIMpulse 7, Nürnberg: Nürnberg Institut for Market Design.

KitaMatch (2023): Startseite, URL: https://kitamatch.com/ (aufgerufen am: 06/08/2025).

Norori, N. / Hu, Q. / Aellen, F. M. / Faraci, F. D. / Tzovara, A. (2021): Addressing Bias in Big Data an AI for Health Care: A Call for Open Science, in: Perspective, Jg. 2 / Nr. 10, Artikel 100347.

Rughinis, R. / Rughinis, C. / Vulpe, S. N. / Rosner, D. (2019): From Social Netizens to Data Citizens: Variations of GDPR awareness in 28 European countries, in: Computer Law & Security Review, Jg. 42, Artikel 105585, DOI: 10.1016/j.clsr.2021.105585.

Smuha, N. A. (2024): The Work of the High-Level Expert Group on AI as the Precursor of the AI Act, in: Ceyhun, N. P. / Forgó, N. / Valcke, P. (Hrsg.): AI Governance and Liability in Europe – A Primer, URL: https://ssrn.com/abstract=5012626 (aufgerufen am: 07/08/2025).

Startup Verband (2024): Startups und Generative KI – Ein neues Zeitalter beginnt, URL: https://startupverband.de/fileadmin/startupverband/forschung/studien/ki/Startups_Generative_KI_2024.pdf (aufgerufen am: 07/08/2025).