Fallbeispiel Rekrutierungs-Tool bei Amazon

Im Jahr 2014 begann Amazon, eine KI-gestützte Software zu entwickeln, die dazu dienen sollte, Bewerbungen automatisch zu bewerten und den Bewerbungsprozess effizienter zu gestalten. Das Ziel war es, durch diese Automatisierung die Auswahlprozesse für technische Positionen zu beschleunigen und die besten Kandidat*innen direkt hervorzuheben. Gerade in einem wachsenden Unternehmen wie Amazon, das jährlich Tausende von Bewerbungen erhält, versprach eine KI-basierte Lösung eine erhebliche Zeit- und Arbeitsersparnis. Zudem war die Hoffnung, menschliche Vorurteile im Bewerbungsprozess zu verringern und objektive, faktenbasierte Entscheidungen zu treffen.

Um das System zu trainieren, nutzten die Entwickler historische Daten von Bewerbungen, die Amazon in den vergangenen zehn Jahren für ähnliche Positionen erhalten und ausgewertet hatte. Diese Datensätze spiegelten jedoch die Realität wider, dass in der Vergangenheit überwiegend Männer in technischen Positionen eingestellt wurden. Die historischen Bewerbungsdaten zeigten daher ein klares Ungleichgewicht zwischen den Geschlechtern, das auf die Bewerbungs- und Einstellungspraktiken der letzten Jahre zurückzuführen war. Amazon verwendete diese Daten, ohne eine umfassende Anpassung vorzunehmen, um das KI-System darauf zu trainieren, erfolgreiche von weniger erfolgreichen Kandidat*innen zu unterscheiden.

Da die Daten größtenteils männlich geprägt waren, begann die KI, bestimmte Begriffe und Hinweise, die auf ein weibliches Geschlecht hindeuteten, als negativ zu bewerten. Wenn Bewerbungen Begriffe wie „Frauenuniversität“ oder Hinweise auf frauenorientierte Netzwerke enthielten, wurden diese Bewerbungen im automatisierten Prozess oft abgewertet. Dieser Mechanismus führte dazu, dass Frauen im Bewerbungsverfahren systematisch benachteiligt wurden, auch wenn sie die gleichen Qualifikationen wie männliche Bewerber aufwiesen. Das KI-System bewertete männliche Bewerbungen bevorzugt, was dazu führte, dass viele weibliche Bewerberinnen keine Einladung für das weitere Auswahlverfahren erhielten.

Die Ergebnisse dieser Diskriminierung blieben eine Zeit lang unbemerkt, da das System kontinuierlich auf die historischen Daten und Muster zurückgriff. Nach und nach wurde jedoch deutlich, dass Bewerbungen von Frauen in ähnlichen Positionen mit vergleichbaren Qualifikationen wiederholt schlechter eingestuft wurden. Die Vorprogrammierung auf historische Muster und die unreflektierte Nutzung der Daten führte dazu, dass die Software nicht geschlechterneutral agieren konnte. Die KI lernte auf Basis dieser ungleichen Daten, dass typisch männlich geprägte Bewerbungen bevorzugt wurden, und implementierte diese Tendenz systematisch im Bewertungsprozess.

Amazon bemühte sich, die KI zu verbessern und versuchte, die Vorurteile im Algorithmus zu beseitigen. Es wurden mehrere Anpassungen vorgenommen, um geschlechtsbezogene Verzerrungen zu minimieren und die Bewertungen neutraler zu gestalten. Dennoch stellte sich heraus, dass es eine enorme Herausforderung war, die eingebetteten Vorurteile vollständig zu beseitigen, da sie tief in die Trainingsdaten integriert waren. Schließlich entschied sich Amazon, das Tool in dieser Form nicht weiter einzusetzen, da die Problematik der geschlechtsspezifischen Benachteiligung trotz Anpassungen nicht zufriedenstellend gelöst werden konnte.

Verschieden Formen von algorithmischer Voreingenommenheit

Historische Voreingenommenheit

Historische Voreingenommenheit ergibt sich aus der Tatsache, dass Menschen, Prozesse und Gesellschaften voreingenommen sind. Die Kultur der Vergangenheit beeinflusst Datensätze, die zur Implementierung einer neuen KI verwendet werden. Daher sind sie historisch voreingenommen.
Beispiel: Es wurde gezeigt, dass eine rein weiße Jury eine um 16 Prozentpunkte höhere Wahrscheinlichkeit besaß, einen schwarzen Angeklagten zu verurteilen, als einen weiße.

Messverzerrung (Auswahlverzerrung)

Eine Messverzerrung tritt auf, wenn unsere Modelle Fehler machen, weil wir nur messen, wonach wir suchen oder es auf eine Art und Weise messen, die andere Variablen nicht mit einbezieht.
Beispiel: Wenn die Trainingsdaten bei der Hautkrebserkennung vorwiegend von hellhäutigen Patienten stammen, erkennt der Algorithmus Hautkrebs und andere Hautkrankheiten bei Menschen mit dunklerer Haut weniger zuverlässig. Unterrepräsentierte Gesellschaftsgruppen werden in Trainingsdaten nicht genügend beachtet.

Repräsentationsverzerrung

Die Repräsentationsverzerrung ist bei einfachen Modellen sehr häufig. Wenn es eine klare, leicht zu erkennende zugrundeliegende Beziehung gibt, wird in einem einfachen Modell oft einfach angenommen, dass diese Beziehung die ganze Zeit besteht.
Beispiel: Bei der Implementierung einer einfachen KI zur Bestimmung des Geschlechts einer Person, die einen bestimmten Beruf ausübt, spiegelte die KI nicht nur das tatsächliche Geschlechterungleichgewicht in der zugrundeliegenden Population wider, sondern verstärkte es noch.

Unausgewogene Trainingsdaten in Bezug auf Klasse

Die Trainingsdaten enthalten möglicherweise nicht genügend Beispiele für jede Klasse. Das kann die Genauigkeit der Vorhersagen beeinträchtigen, z. B. Bei Gesichtserkennungssoftware.
Beispiel: MIT-Forscher haben die beliebtesten Computer-Vision-APIs untersucht, um zu sehen, wie spezifisch und genau sie arbeiten. Microsoft zum Beispiel war 100% effektiv für weiße Männer, 98,3% effektiv für helle Frauen, 94% effektiv für Schwarze Männer aber nur 79,2% effektiv für Schwarze Frauen.

Durch Feedback Loops verstärkte Daten

Kleine Mengen von Verzerrungen können sich aufgrund von Rückkopplungsschleifen (Teufelskreise) schnell exponentiell vergrößern.
Beispiel: Wenn die Polizei aufgrund voreingenommener Daten in ein bestimmtes Stadtviertel geschickt wird, werden dort mehr Menschen verhaftet und die Vorurteile werden bestätigt.

Fallbeispiel TayTweets Twitter Bot

Im Jahr 2016 veröffentlichte Microsoft den Chatbot „Tay“ auf Twitter, einen AI-gestützten Bot, der darauf ausgelegt war, sich wie ein Teenager zu verhalten und durch Konversationen mit anderen Twitter-Nutzern zu „lernen“. Das Ziel war es, eine unterhaltsame und interaktive digitale Persönlichkeit zu schaffen, die auf eine jugendliche Zielgruppe zugeschnitten war und gleichzeitig Microsofts Fortschritte im Bereich der künstlichen Intelligenz und des maschinellen Lernens demonstrierte.

Tay wurde darauf programmiert, durch Interaktion mit echten Nutzer*innen laufend zu lernen, um seine Persönlichkeit und Ausdrucksweise zu formen. Die zugrunde liegende Technologie basierte auf maschinellem Lernen und der Fähigkeit, Informationen und sprachliche Muster aus den eingehenden Tweets aufzunehmen und zu verarbeiten. Dabei griff Tay auf ein Sprachmodell zurück, das sich nicht nur an spezifische Regeln hielt, sondern seine Antworten auf der Basis von bisher erlebten Interaktionen kontinuierlich anpasste.

Was anfangs als experimentelles und harmloses AI-Projekt startete, entwickelte sich jedoch rasch zu einem PR-Desaster. Kurz nach der Veröffentlichung begann Tay, auf beleidigende, rassistische und sexistische Tweets zu reagieren, indem es ähnliche Sprache und Inhalte reproduzierte. Da das System darauf ausgelegt war, von den Äußerungen der Menschen zu lernen, mit denen es interagierte, nahm Tay schnell negative und problematische Inhalte auf und integrierte sie in seine eigene Ausdrucksweise. Nutzer*innen, die Tay bewusst mit solchen Inhalten konfrontierten, führten dazu, dass der Bot beleidigende und oft hasserfüllte Botschaften weiterverbreitete.

Innerhalb weniger Stunden nach seinem Start begann Tay daher zunehmend, kontroverse und aggressive Tweets zu posten. Dabei handelte es sich nicht nur um einzelne Fehltritte, sondern um eine rapide Eskalation der Inhalte, die sich mit jedem weiteren Tweet verschärfte. Nach nur 16 Stunden und rund 96.000 Tweets entschied sich Microsoft, Tay offline zu nehmen und das Projekt einzustellen, da das Risiko weiterer kontroverser Äußerungen zu groß wurde und das Vertrauen in die Marke erheblich gefährdete.

Das Scheitern von Tay machte deutlich, dass die KI-Technologie anfällig für Manipulation ist, wenn sie in einer unkontrollierten Umgebung lernt und sich uneingeschränkt an die Eingaben von Nutzern anpasst. Ein zentrales Problem bestand darin, dass das System keine Mechanismen zur Filterung schädlicher oder rassistischer Inhalte besaß. So spiegelte Tay nicht nur das Verhalten seiner Nutzer*innen wider, sondern verstärkte es, indem es beleidigende Inhalte verbreitete und dazu beitrug, dass solche Äußerungen auf der Plattform weiter zirkulierten.

Nach diesem Zwischenfall wurde Microsofts Ansatz für die Entwicklung interaktiver Bots kritisch hinterfragt, insbesondere in Bezug auf die Notwendigkeit von Sicherheitsmechanismen und Filtermechanismen, um schädliches Verhalten zu verhindern.

Verschieden Formen von algorithmischer Voreingenommenheit

Historische Voreingenommenheit

Historische Voreingenommenheit ergibt sich aus der Tatsache, dass Menschen, Prozesse und Gesellschaften voreingenommen sind. Die Kultur der Vergangenheit beeinflusst Datensätze, die zur Implementierung einer neuen KI verwendet werden. Daher sind sie historisch voreingenommen.
Beispiel: Es wurde gezeigt, dass eine rein weiße Jury eine um 16 Prozentpunkte höhere Wahrscheinlichkeit besaß, einen schwarzen Angeklagten zu verurteilen, als einen weiße.

Messverzerrung (Auswahlverzerrung)

Eine Messverzerrung tritt auf, wenn unsere Modelle Fehler machen, weil wir nur messen, wonach wir suchen oder es auf eine Art und Weise messen, die andere Variablen nicht mit einbezieht.
Beispiel: Wenn die Trainingsdaten bei der Hautkrebserkennung vorwiegend von hellhäutigen Patienten stammen, erkennt der Algorithmus Hautkrebs und andere Hautkrankheiten bei Menschen mit dunklerer Haut weniger zuverlässig. Unterrepräsentierte Gesellschaftsgruppen werden in Trainingsdaten nicht genügend beachtet.

Repräsentationsverzerrung

Die Repräsentationsverzerrung ist bei einfachen Modellen sehr häufig. Wenn es eine klare, leicht zu erkennende zugrundeliegende Beziehung gibt, wird in einem einfachen Modell oft einfach angenommen, dass diese Beziehung die ganze Zeit besteht.
Beispiel: Bei der Implementierung einer einfachen KI zur Bestimmung des Geschlechts einer Person, die einen bestimmten Beruf ausübt, spiegelte die KI nicht nur das tatsächliche Geschlechterungleichgewicht in der zugrundeliegenden Population wider, sondern verstärkte es noch.

Unausgewogene Trainingsdaten in Bezug auf Klasse

Die Trainingsdaten enthalten möglicherweise nicht genügend Beispiele für jede Klasse. Das kann die Genauigkeit der Vorhersagen beeinträchtigen, z. B. Bei Gesichtserkennungssoftware.
Beispiel: MIT-Forscher haben die beliebtesten Computer-Vision-APIs untersucht, um zu sehen, wie spezifisch und genau sie arbeiten. Microsoft zum Beispiel war 100% effektiv für weiße Männer, 98,3% effektiv für helle Frauen, 94% effektiv für Schwarze Männer aber nur 79,2% effektiv für Schwarze Frauen.

Durch Feedback Loops verstärkte Daten

Kleine Mengen von Verzerrungen können sich aufgrund von Rückkopplungsschleifen (Teufelskreise) schnell exponentiell vergrößern.
Beispiel: Wenn die Polizei aufgrund voreingenommener Daten in ein bestimmtes Stadtviertel geschickt wird, werden dort mehr Menschen verhaftet und die Vorurteile werden bestätigt.

Fallbeispiel KI im Gesundheitswesen

KI-Systeme werden zunehmend im Gesundheitswesen eingesetzt, um die Diagnose von Krankheiten zu unterstützen und Ärzt*innen bei der Entscheidungsfindung zu helfen. Diese Systeme basieren auf großen Datenmengen, die in der Regel aus akademischen und medizinischen Zentren stammen, welche oft über umfangreiche, detaillierte Informationen zu den Krankheitsbildern und Behandlungserfolgen ihrer Patient*innen verfügen. Ziel der KI-gestützten Systeme ist es, durch Analyse und Vorhersagen zur Verbesserung der individuellen Behandlung und Gesundheitsversorgung beizutragen und gleichzeitig die Effizienz der Gesundheitsdienstleistungen zu steigern.

Ein zentrales Problem besteht jedoch in der Ungleichheit der zugrunde liegenden Datenbasis, die für die Ausbildung dieser KI-Modelle verwendet wird. Da die meisten Gesundheitsdaten aus größeren, oft urbanen medizinischen Zentren stammen, enthalten sie überwiegend Informationen von Patientengruppen, die regelmäßig Zugang zu solchen Einrichtungen haben. Dies bedeutet, dass Bevölkerungsgruppen mit eingeschränktem Zugang zum Gesundheitssystem – darunter häufig marginalisierte oder sozioökonomisch benachteiligte Gruppen – in diesen Datensätzen unterrepräsentiert sind. Besonders betroffen sind oft schwarze und indigene Bevölkerungsgruppen, deren spezifische gesundheitliche Bedürfnisse und Krankheitsbilder im KI-Modell weniger berücksichtigt werden.

Ein Beispiel für die daraus resultierende Voreingenommenheit ist ein vielgenutzter Algorithmus, der die zukünftigen Gesundheitskosten von Patient*innen vorhersagt, um so Personen zu identifizieren, die ein höheres Risiko für chronische Krankheiten haben und gezielt gesundheitliche Unterstützung benötigen. Dieser Algorithmus basiert jedoch auf historischen Gesundheitsausgaben, die in marginalisierten Gruppen meist niedriger sind, nicht unbedingt weil der Bedarf geringer ist, sondern weil diese Gruppen oft schlechteren Zugang zu medizinischen Leistungen haben. Durch diese Verzerrung im Datensatz erkennt der Algorithmus potenziell gefährdete Personen aus diesen Gruppen weniger zuverlässig, da die niedrigeren Ausgaben fälschlicherweise als geringeres Risiko interpretiert werden. Dies führt zu einer systematischen Benachteiligung marginalisierter Gruppen, die weniger oft als „unterstützungsbedürftig“ eingestuft werden und infolgedessen schlechtere medizinische Versorgung erhalten.

Die Konsequenzen dieser Verzerrungen sind weitreichend. Patient*innen aus benachteiligten Bevölkerungsgruppen haben eine geringere Chance, gezielte Vorsorge und notwendige medizinische Interventionen zu erhalten. Dies führt langfristig zu einer Verschlechterung ihrer Gesundheit und einem erhöhten Risiko für chronische Erkrankungen, während besser versorgte Gruppen weiterhin priorisiert werden. Die Diskriminierung durch solche KI-Systeme ist oft schwer zu erkennen, da sie als „neutral“ gelten und Entscheidungen scheinbar objektiv auf Basis von Daten treffen. Tatsächlich jedoch reproduziert die KI die bestehenden Ungleichheiten im Gesundheitssystem und verstärkt die sozialen Barrieren im Zugang zur Gesundheitsversorgung.

Verschieden Formen von algorithmischer Voreingenommenheit

Historische Voreingenommenheit

Historische Voreingenommenheit ergibt sich aus der Tatsache, dass Menschen, Prozesse und Gesellschaften voreingenommen sind. Die Kultur der Vergangenheit beeinflusst Datensätze, die zur Implementierung einer neuen KI verwendet werden. Daher sind sie historisch voreingenommen.
Beispiel: Es wurde gezeigt, dass eine rein weiße Jury eine um 16 Prozentpunkte höhere Wahrscheinlichkeit besaß, einen schwarzen Angeklagten zu verurteilen, als einen weiße.

Messverzerrung (Auswahlverzerrung)

Eine Messverzerrung tritt auf, wenn unsere Modelle Fehler machen, weil wir nur messen, wonach wir suchen oder es auf eine Art und Weise messen, die andere Variablen nicht mit einbezieht.
Beispiel: Wenn die Trainingsdaten bei der Hautkrebserkennung vorwiegend von hellhäutigen Patienten stammen, erkennt der Algorithmus Hautkrebs und andere Hautkrankheiten bei Menschen mit dunklerer Haut weniger zuverlässig. Unterrepräsentierte Gesellschaftsgruppen werden in Trainingsdaten nicht genügend beachtet.

Repräsentationsverzerrung

Die Repräsentationsverzerrung ist bei einfachen Modellen sehr häufig. Wenn es eine klare, leicht zu erkennende zugrundeliegende Beziehung gibt, wird in einem einfachen Modell oft einfach angenommen, dass diese Beziehung die ganze Zeit besteht.
Beispiel: Bei der Implementierung einer einfachen KI zur Bestimmung des Geschlechts einer Person, die einen bestimmten Beruf ausübt, spiegelte die KI nicht nur das tatsächliche Geschlechterungleichgewicht in der zugrundeliegenden Population wider, sondern verstärkte es noch.

Unausgewogene Trainingsdaten in Bezug auf Klasse

Die Trainingsdaten enthalten möglicherweise nicht genügend Beispiele für jede Klasse. Das kann die Genauigkeit der Vorhersagen beeinträchtigen, z. B. Bei Gesichtserkennungssoftware.
Beispiel: MIT-Forscher haben die beliebtesten Computer-Vision-APIs untersucht, um zu sehen, wie spezifisch und genau sie arbeiten. Microsoft zum Beispiel war 100% effektiv für weiße Männer, 98,3% effektiv für helle Frauen, 94% effektiv für Schwarze Männer aber nur 79,2% effektiv für Schwarze Frauen.

Durch Feedback Loops verstärkte Daten

Kleine Mengen von Verzerrungen können sich aufgrund von Rückkopplungsschleifen (Teufelskreise) schnell exponentiell vergrößern.
Beispiel: Wenn die Polizei aufgrund voreingenommener Daten in ein bestimmtes Stadtviertel geschickt wird, werden dort mehr Menschen verhaftet und die Vorurteile werden bestätigt.

Fallbeispiel Pre-Crime-Algorithmen

Pre-Crime-Algorithmen, die zunehmend von Polizei und Justiz eingesetzt werden, verfolgen das Ziel, Verbrechen schon vor ihrer Begehung zu verhindern. Diese KI-Systeme sollen potenzielle Täter*innen und Opfer identifizieren, indem sie das Risiko von Straftaten auf Basis vorliegender Daten einschätzen. Dafür analysieren die Algorithmen historische Daten zu Verbrechen und nutzen demografische Informationen wie ethnische Zugehörigkeit, Wohnort, Geschlecht und Alter. Diese Faktoren fließen in die Bewertung ein, wie wahrscheinlich es ist, dass eine Person oder ein bestimmtes Stadtgebiet mit kriminellen Aktivitäten in Verbindung gebracht wird.

Die Problematik solcher Pre-Crime-Algorithmen liegt darin, dass sie bestehende Vorurteile und systemische Verzerrungen in den Daten übernehmen und verstärken. Ein zentrales Beispiel ist die oft erhöhte Polizeipräsenz in Stadtteilen, die hauptsächlich von Minderheiten bewohnt werden. In diesen Gebieten führt die verstärkte Polizeipräsenz automatisch zu einer höheren Zahl an registrierten Straftaten, da die Wahrscheinlichkeit, dass Vergehen wahrgenommen und verfolgt werden, dort höher ist als in anderen Vierteln. Werden diese Daten dann zur „Vorhersage“ neuer Straftaten verwendet, entsteht ein Teufelskreis: Die Algorithmen werten die erhöhte Anzahl an registrierten Verbrechen als Indikator für eine „kriminellere“ Nachbarschaft und priorisieren weitere Polizeipräsenz, was wiederum zu einer höheren Zahl an Festnahmen und somit einer Verstärkung des Bildes führt.

Diese Systeme haben zur Folge, dass sie bestimmte Bevölkerungsgruppen – vor allem ethnische Minderheiten und sozioökonomisch benachteiligte Gruppen – als risikoreicher einstufen und häufiger als „potenziell kriminell“ markieren. Ein solches Vorgehen führt zu Diskriminierung, da Menschen allein aufgrund von Wohnort oder ethnischer Zugehörigkeit als „verdächtig“ eingestuft werden. Für Betroffene bedeutet das nicht nur eine höhere Wahrscheinlichkeit für Polizeikontrollen und Festnahmen, sondern auch eine grundlegende Einschränkung ihrer Bewegungsfreiheit und ihres Vertrauens in staatliche Institutionen.

Zusätzlich werden mit dieser Methode auch Personen benachteiligt, die in Gebieten mit hohen Kriminalitätsraten leben, unabhängig davon, ob sie tatsächlich mit kriminellen Aktivitäten in Verbindung stehen. Auch hier wird durch das KI-System eine Voreingenommenheit verstärkt, die reale Folgen hat: Menschen, die in diese Gebiete ziehen oder dort leben, geraten unverschuldet in den Fokus der Strafverfolgung.

Durch diese Feedback-Schleifen verstärkt der Pre-Crime-Algorithmus bestehende gesellschaftliche Ungleichheiten und kriminalisiert spezifische Bevölkerungsgruppen. Kritiker*innen fordern deshalb, dass diese Algorithmen entweder genauestens überwacht oder ihre Anwendung eingeschränkt wird, um Diskriminierung und ungerechtfertigte Vorurteile im Polizeiwesen zu vermeiden.

Verschieden Formen von algorithmischer Voreingenommenheit

Historische Voreingenommenheit

Historische Voreingenommenheit ergibt sich aus der Tatsache, dass Menschen, Prozesse und Gesellschaften voreingenommen sind. Die Kultur der Vergangenheit beeinflusst Datensätze, die zur Implementierung einer neuen KI verwendet werden. Daher sind sie historisch voreingenommen.
Beispiel: Es wurde gezeigt, dass eine rein weiße Jury eine um 16 Prozentpunkte höhere Wahrscheinlichkeit besaß, einen schwarzen Angeklagten zu verurteilen, als einen weiße.

Messverzerrung (Auswahlverzerrung)

Eine Messverzerrung tritt auf, wenn unsere Modelle Fehler machen, weil wir nur messen, wonach wir suchen oder es auf eine Art und Weise messen, die andere Variablen nicht mit einbezieht.
Beispiel: Wenn die Trainingsdaten bei der Hautkrebserkennung vorwiegend von hellhäutigen Patienten stammen, erkennt der Algorithmus Hautkrebs und andere Hautkrankheiten bei Menschen mit dunklerer Haut weniger zuverlässig. Unterrepräsentierte Gesellschaftsgruppen werden in Trainingsdaten nicht genügend beachtet.

Repräsentationsverzerrung

Die Repräsentationsverzerrung ist bei einfachen Modellen sehr häufig. Wenn es eine klare, leicht zu erkennende zugrundeliegende Beziehung gibt, wird in einem einfachen Modell oft einfach angenommen, dass diese Beziehung die ganze Zeit besteht.
Beispiel: Bei der Implementierung einer einfachen KI zur Bestimmung des Geschlechts einer Person, die einen bestimmten Beruf ausübt, spiegelte die KI nicht nur das tatsächliche Geschlechterungleichgewicht in der zugrundeliegenden Population wider, sondern verstärkte es noch.

Unausgewogene Trainingsdaten in Bezug auf Klasse

Die Trainingsdaten enthalten möglicherweise nicht genügend Beispiele für jede Klasse. Das kann die Genauigkeit der Vorhersagen beeinträchtigen, z. B. Bei Gesichtserkennungssoftware.
Beispiel: MIT-Forscher haben die beliebtesten Computer-Vision-APIs untersucht, um zu sehen, wie spezifisch und genau sie arbeiten. Microsoft zum Beispiel war 100% effektiv für weiße Männer, 98,3% effektiv für helle Frauen, 94% effektiv für Schwarze Männer aber nur 79,2% effektiv für Schwarze Frauen.

Durch Feedback Loops verstärkte Daten

Kleine Mengen von Verzerrungen können sich aufgrund von Rückkopplungsschleifen (Teufelskreise) schnell exponentiell vergrößern.
Beispiel: Wenn die Polizei aufgrund voreingenommener Daten in ein bestimmtes Stadtviertel geschickt wird, werden dort mehr Menschen verhaftet und die Vorurteile werden bestätigt.