

Zeitschrift **Frauenfragen**
Revue **Questions au féminin**
Rivista **Questioni femminili**



KI, Algorithmen und Geschlecht IA, algorithmes et genre IA, algoritmi e genere



Schweizerische Eidgenossenschaft
Confédération suisse
Confederazione Svizzera
Confederaziun svizra

Eidgenössische Kommission für Frauenfragen EKF
Commission fédérale pour les questions féminines CFQF
Commissione federale per le questioni femminili CFQF

Die Eidgenössische Kommission für Frauenfragen EKF ist eine ausserparlamentarische Kommission des Bundes. Sie analysiert die Situation der Frauen in der Schweiz und setzt sich für die Gleichstellung der Geschlechter ein.

La Commission fédérale pour les questions féminines CFQF est une commission extraparlamentaire de la Confédération. Elle analyse la condition des femmes en Suisse et travaille en faveur de l'égalité des sexes.

La Commissione federale per le questioni femminili CFQF è una commissione extraparlamentare della Confederazione. Essa analizza la situazione delle donne in Svizzera e si impegna a promuovere la parità tra i sessi.

- 4 **Bettina Fredrich | Caroline Honegger**
- 5 Editorial
- 5 Éditorial
- 6 Editoriale

- 7 **Alizé Rose-May Monod**
- 7 Illustrationen: Im Gespräch mit der KI
- 8 Illustrations : Conversation avec une IA
- 9 Illustrazioni: A colloquio con un'IA

Was bedeuten KI und Algorithmen für die Gleichstellung?
L'IA et les algorithmes, quelle signification pour l'égalité ?
Cosa significano l'IA e gli algoritmi per l'uguaglianza?

- 12 **Bianca Prietl**
Universität Basel, Zentrum Gender Studies
Hat Künstliche Intelligenz ein Geschlecht?
- 16 **Yara Abu Awad**
Office fédéral de la statistique OFS
« Les algorithmes ont un impact direct sur nos vies, il est donc essentiel qu'ils soient équitables »
- 20 **Mascha Kurpicz-Briki**
Bernern Fachhochschule, Departement Technik und Informatik
Gesellschaftliche Stereotype in Sprachmodellen: Herausforderungen und Stand der Forschung
- 26 **Angela Müller**
AlgorithmWatch CH
«Es geht nicht darum, Technologie zu verteufeln»
- 30 **Monica Landoni**
Università della Svizzera italiana, facoltà di Scienze Informatiche
TADAA – ovvero come aumentare la consapevolezza dell'esistenza di stereotipi di genere in informatica

- Deutsch
- Français
- Italiano
- English

Chancen und Herausforderungen
Défis et opportunités
Opportunità e sfide

- 36 **David Raedler**
Avocat spécialisé en droit du travail
«Il est préférable que des humains contrôlent le travail de l'IA»
- 39 **Annabella Bassler**
EqualVoice-Initiative von Ringier
«Künstliche Intelligenz hat das Potenzial, Geschlechterungleichheiten zu minimieren»
- 44 **Cristina Rossi**
Impresa di senologia b-rayZ
«L'intelligenza artificiale è in grado di analizzare una mammografia come lo farebbe un essere umano»
- 48 **Morgane Bonvallat**
Projet Stop Hate Speech de la Public Discourse Foundation
Bot Dog: une intelligence artificielle au secours de la justice sociale
- 54 **Laurent Sedano**
Medienbildner und Digitalisierungsexperte
«Nicht die KI ist rassistisch oder sexistisch – wir sind es»
- 58 **Jolanda Suter**
Illustrationen: Es wimmelt
Illustrations: Ça foisonne
Illustrazioni: Immagini brulicanti

Rechtliche Grundlagen und politische Diskussionen
Bases légales et débats politiques
Basi legali e dibattiti politici

- 62 **Fabian Lütz**
Université de Lausanne, faculté de droit
La régulation de l'IA en Suisse et en Europe
- 68 **Ivana Bartoletti**
Esperta di IA e diritti di genere presso il Consiglio d'Europa
«La discriminazione degli algoritmi non è solo una questione tecnica, ma è anche sociale e politica»
- 72 **Caitlin Kraft-Buchman**
Fondatrice du think tank Women at the Table
«Nous devons assumer la responsabilité des préjugés qui sont au cœur des modèles algorithmiques»
- 78 **Liliane Obrecht**
Universität Basel, Juristische Fakultät
«Öffentliche Verzeichnisse sind ideal, um über die Risiken und Chancen von KI zu diskutieren»
- 82 **Min Li Marti | Franz Grüter**
Parlamentarische Gruppe Parldigi
Wo steht die Schweizer Politik im Bereich KI? Das Co-Präsidium von Parldigi liefert Antworten

Illustrationen
Illustrations
Illustrazioni

- **Alizé Rose-May Monod**
- 10 Innovation, Revolution, Uranus
- 24 Hyenas Hunting
- 34 #AIDemonstration2099
- 42 Asylum, Migration, and AI
- 52 Bot Dog
- 66 We exist
- 76 Swiss Money
- 86 Feline Surveillance

Editorial

Bettina Fredrich | Caroline Honegger



Von Video-Empfehlungen auf Netflix zur personalisierten Werbung im Internetbrowser: Algorithmen haben längst Einzug in unseren Alltag gefunden und beeinflussen diesen massgeblich. Mit der Lancierung von generativer Künstlicher Intelligenz, beispielsweise in Form von Chatbots und Bildprogrammen, hat die Entwicklung an Tempo und Reichweite zugelegt. Die Auswirkungen der neuen Technologien können wir erst ansatzweise erahnen.

Denn algorithmische Systeme und Künstliche Intelligenz sind weder neutral noch objektiv, sondern stützen sich auf vorhandene Daten, welche vielfach gesellschaftliche Ungleichheiten beinhalten und diese somit reproduzieren. Aus einer Gleichstellungsperspektive stellt sich die Frage, welche technischen, politischen, rechtlichen und gesellschaftlichen Anpassungen nötig sind, um sicherzustellen, dass diese neue Technologie so eingesetzt wird, dass sie die Geschlechtergleichstellung fördert und ihr nicht zuwiderläuft.

Was braucht es für eine geschlechtergerechte KI?

Das vorliegende Heft beleuchtet das Thema Künstliche Intelligenz und Algorithmen aus einer Geschlechterperspektive. Wir befassen uns mit technischen Fragen: Wie funktionieren Algorithmen konkret und welche Möglichkeiten gibt es, sie so zu programmieren, dass sie vorurteilsfrei funktionieren? Wir werfen einen Blick auf rechtliche und politische Aspekte: Welchen juristischen Rahmen braucht es in der Schweiz, um vor Diskriminierung durch KI zu schützen, und welche Fragen stellen sich auf europäischer Ebene? Ausserdem möchten wir aufzeigen, dass Künstliche Intelligenz und Algorithmen sowohl Risiko als auch Chance für die Gleichstellung sein können: Konkrete Beispiele verdeutlichen, wie Algorithmen im Alltag diskriminieren – oder zu einer gerechteren Gesellschaft beitragen. Mit Fachartikeln, Interviews und Porträts beleuchten wir die aktuelle Situation in der Schweiz.

Illustriert wird das Heft von zwei Kunstschaffenden, die sich auf ganz unterschiedliche Weise mit dem Thema KI auseinandersetzen. Alizé Rose-May Monods Bilder sind aus einem angeregten Dialog mit dem KI-Tool DALL-E entstanden und Jolanda Suter stellt den digitalen Alltag in drei humorvollen Wimmelbildern dar.

Erweiterte Perspektive

Während die Frage der Unterrepräsentation und Diskriminierung von Frauen im Kontext der Digitalisierung weiterhin relevant ist, wollen wir im vorliegenden Heft auch eine breitere Geschlechterdebatte einschliessen und abbilden. Nonbinäre Personen bleiben in vielen Datensätzen unsichtbar und queere Menschen sind besonders häufig Zielscheibe von Online-Hass. Wir haben die Autor*innen im Heft deshalb aufgefordert, eine umfassende Geschlechterperspektive einzunehmen. Dies beinhaltet auch, dass sie alle ihre bevorzugte Art der geschlechtergerechten Sprache verwenden. Die sprachliche Vielfalt spiegelt die gesellschaftliche Vielfalt wider.

Die EKF wünscht Ihnen eine anregende Lektüre.

Bettina Fredrich, Geschäftsleiterin EKF bis August 2024
Caroline Honegger, wissenschaftliche Mitarbeiterin EKF

«Algorithmische Systeme
und Künstliche Intelligenz
sind weder neutral noch
objektiv.»

Éditorial

Bettina Fredrich | Caroline Honegger



De simples recommandations de vidéos sur Netflix à la publicité personnalisée sur les navigateurs : cela fait longtemps que les algorithmes se sont installés dans notre quotidien, sur lequel ils ont une influence non négligeable. Cette évolution a pris de l'ampleur et de la vitesse avec le lancement de l'intelligence artificielle générative, par exemple sous la forme de chatbots ou de logiciels de génération d'images. Avec quelles répercussions ? Difficile de le prévoir pour le moment.

Les systèmes algorithmiques et l'intelligence artificielle ne sont ni neutres ni objectifs. Ils reposent sur des données existantes, truffées de discriminations variées qu'ils reproduisent. Du point de vue de l'égalité, cela interroge sur la nécessité de procéder à des adaptations techniques, politiques, juridiques et sociales afin de s'assurer que cette nouvelle technologie soit employée de manière à améliorer l'égalité des genres, et non pas à l'entraver.

Que faut-il pour une IA équitable du point de vue du genre ?

La présente édition de notre revue présente le sujet de l'intelligence artificielle et des algorithmes sous l'angle du genre. Nous abordons des questions techniques : comment les algorithmes fonctionnent-ils concrètement ? Comment peut-on les programmer de façon à ce qu'ils ne reproduisent pas de préjugés ? Nous envisageons aussi les aspects juridiques et politiques : quel cadre légal faut-il en Suisse pour protéger contre la discrimination par l'IA ? Quelles sont les questions qui se posent au niveau européen ? Nous nous employons en outre à montrer que l'intelligence artificielle et les algorithmes peuvent être à la fois un risque et une opportunité pour l'égalité : des exemples concrets illustrent comment des algorithmes produisent des discriminations dans la vie quotidienne ... ou comment ils contribuent à une société plus égalitaire. Des articles, des interviews et des portraits mettent en lumière la situation actuelle en Suisse.

« Les systèmes algorithmiques et l'intelligence artificielle ne sont ni neutres ni objectifs. »

Ce numéro est illustré par deux artistes, qui abordent la question de l'IA de manière très différente. Les créations d'Alizé Rose-May Monod sont le fruit d'un vif dialogue avec l'outil d'IA DALL-E tandis que Jolanda Suter décrit notre quotidien numérique à travers trois illustrations pleines d'humour.

Une perspective élargie

Si la question de la sous-représentation et de la discrimination des femmes dans le contexte de la transformation numérique reste pertinente, nous souhaitons aussi faire place à un débat plus large sur le genre. Les personnes non binaires restent invisibles dans de nombreuses données, et les personnes queer sont particulièrement souvent la cible de phénomènes haineux en ligne. Nous avons donc invité les auteur*es qui contribuent à cette édition à adopter un point de vue global sur le genre et à utiliser l'écriture épïcène de leur choix. La diversité langagière reflète la diversité de la société.

La CFQF vous souhaite une lecture enrichissante.

Bettina Fredrich, directrice de la CFQF jusqu'en août 2024

Caroline Honegger, collaboratrice scientifique de la CFQF

Editoriale

Bettina Fredrich | Caroline Honegger



Dalle raccomandazioni video su Netflix alla pubblicità personalizzata sul navigatore web, gli algoritmi sono entrati da tempo nella nostra vita quotidiana e la influenzano fortemente. Il lancio dell'intelligenza artificiale generativa, per esempio sotto forma di chatbot e programmi generatori d'immagini, ha aumentato la velocità e la portata di questo sviluppo tecnologico, i cui impatti sono oggi difficili da prevedere.

I sistemi algoritmici e l'intelligenza artificiale, infatti, non sono né neutrali né obiettivi, ma si basano sui dati disponibili che spesso contengono disuguaglianze sociali che vengono riprodotte. Quali adeguamenti tecnici, politici, giuridici e sociali sono allora necessari per fare in modo che questa nuova tecnologia sia impiegata per migliorare la parità di genere, e non per contrastarla?

Cosa occorre per un'IA rispettosa del genere?

Il presente numero di «Questioni femminili» ha come tema l'intelligenza artificiale e gli algoritmi in una prospettiva di genere. Abordiamo questioni tecniche: come funzionano concretamente gli algoritmi? Come possono essere programmati affinché siano liberi da pregiudizi? Gettiamo anche uno sguardo sugli aspetti politico-giuridici chiarendo il quadro legale necessario in Svizzera per proteggere dalla discriminazione a opera dell'IA e presentando le questioni che si pongono a livello europeo. Con esempi concreti mostriamo come l'intelligenza artificiale e gli algoritmi possano costituire sia un rischio sia un'opportunità per l'uguaglianza. Articoli specialistici, interviste e ritratti fanno luce sulla situazione attuale in Svizzera.

«I sistemi algoritmici e l'intelligenza artificiale non sono né neutrali né obiettivi.»

La rivista è illustrata da due artiste* che affrontano il tema dell'IA in modi molto diversi. Le immagini di Alizé Rose-May Monod sono il risultato di un dialogo vivace con lo strumento IA DALL-E, mentre Jolanda Suter ritrae la vita digitale di tutti i giorni in tre umoristiche illustrazioni.

Una prospettiva allargata

Sebbene la questione della sottorappresentanza e della discriminazione delle donne nel contesto della digitalizzazione rimanga rilevante, in questa rivista vogliamo includere e sviluppare anche un dibattito più ampio sul tema del genere. Le persone non binarie rimangono invisibili in numerosi set di dati e quelle queer sono molto spesso bersaglio di odio online. Abbiamo quindi esortato le autrici e gli autori* ad adottare una prospettiva di genere a 360 gradi e a utilizzare il loro linguaggio inclusivo di predilezione. La diversità linguistica riflette la diversità sociale.

La CFQF vi augura una lettura avvincente.

Bettina Fredrich, responsabile del Segretariato della CFQF fino ad agosto 2024
Caroline Honegger, collaboratrice scientifica della CFQF

Illustrationen: Im Gespräch mit der KI

Dieses Heft «Frauenfragen» wird unter anderem von **Alizé Rose-May Monod** (ARMM) illustriert. ARMM ist Künstler*in und Kulturschaffende* und lebt seit bald zehn Jahren in Bern. ARMMs transdisziplinäre Arbeit kreist um die Themen Beziehungen und Kollektivität und spiegelt ARMMs vielfältiges Engagement wider. Dazu gehören Kuratierung, Lehre, Recherche und Aktivismus ebenso wie die Arbeit mit Textil, Video, Text und Objekt. ARMM nutzt feministische Theorie als Basis für die künstlerische Praxis und versteht Queerness – auch die eigene – als Linse und Arbeitsmethode. So schafft ARMM Kunst, die über hegemoniale Handlungs- und Seinsweisen hinausgeht und Alternativen aufzeigt.

Für die Zeitschrift «Frauenfragen» arbeitet ARMM erstmals mit einem KI-Bildgenerator. Dabei ist ein Dialog mit der Künstlichen Intelligenz entstanden – immer aus einer feministischen Perspektive.

DALL-E und ich: ein künstlerisches Experiment

Zu ihrem Vorgehen schreibt ARMM: «Die Eidgenössische Kommission für Frauenfragen EKF lud mich ein, die aktuelle Ausgabe der Zeitschrift «Frauenfragen», die sich mit der Nutzung von Künstlicher Intelligenz und Algorithmen aus einer Geschlechterperspektive befasst, zu illustrieren. Mir war schnell klar, auf diesen Auftrag mit der Nutzung einer Künstlichen Intelligenz zu antworten: DALL-E.

Da ich die Arbeit im Kollektiv gewöhnt bin, sah ich in der Kooperation mit einer KI die Möglichkeit, diese Praxis auf eine neue Art zu erweitern. Für jede Illustration legte ich den Inhalt, die symbolischen Bezüge und den gewünschten Stil fest und bat DALL-E, ein Bild zu entwerfen. In diesem Prozess ist mein Beitrag entscheidend, da ich die KI selbst «füttere»: Sie wendet an, was ich ihr gebe, solange es nicht gegen die Nutzungs- und Inhaltsregeln verstösst. Dass das relativ schnell der Fall ist, wurde mir klar, als ich um ein Bild von zwei sich küssenden Lesben oder einer Schweizer Banknote bat.

Vor Ihnen liegt nun das visuelle Ergebnis meiner Konversation mit DALL-E – ein Hin und Her aus Fragen, Austausch, Neujustierungen und Staunen. Jedes Bild wird von einer textlichen Zusammenfassung dieses «Gesprächs» begleitet, das manchmal über die visuelle Produktion hinausging und hin und wieder nahe am Aktivismus war. Wenn Sie sich wundern über das, was Sie sehen oder lesen, denken Sie daran: Alles ist das Ergebnis eines kontextabhängigen Blicks – die Künstliche Intelligenz mit eingeschlossen.»

Illustrations : Conversation avec une IA

La présente revue «Questions au féminin» est illustrée, entre autres, par **Alizé Rose-May Monod** (ARMM). ARMM est artiste et travailleur*euse culturel*le. Son travail transdisciplinaire tourne autour de la thématique des relations et de la collectivité. Il reflète la diversité de son engagement : curation, enseignement, recherche et activisme, mais aussi création utilisant le textile, la vidéo, le texte et l'objet. Les théories féministes sont à la base de la pratique artistique d'ARMM, pour qui la *queerness* – également la sienne – est à la fois un prisme et une méthode de travail. ARMM crée un art qui transcende les manières hégémoniques d'être et d'agir pour ouvrir d'autres voies.

Pour «Questions au féminin», ARMM a travaillé pour la première fois avec un générateur d'images par IA. Un dialogue avec l'intelligence artificielle s'est ainsi engagé, toujours dans une perspective féministe.

DALL-E et moi : une expérience artistique

Concernant sa démarche, l'artiste écrit : «La Commission fédérale pour les questions féminines CFQF m'a invité*e à illustrer la nouvelle édition de sa revue «Questions au féminin», consacrée à l'utilisation de l'intelligence artificielle et des algorithmes dans une perspective de genre. Il m'est rapidement apparu comme une évidence que j'allais recourir à une IA : DALL-E.

Habitué*e à travailler en collectif, j'ai vu dans la coopération avec une IA la possibilité d'élargir cette pratique. Pour chaque illustration, j'ai défini le contenu, les références symboliques et le style que je souhaitais, puis j'ai demandé à DALL-E de générer une image. Dans cette démarche, mon apport est déterminant car c'est moi qui «nourris» l'IA : elle utilise ce que je lui donne tant que cela ne contrevient pas aux règles d'utilisation et de contenu. J'ai compris que l'on atteignait vite ces limites lorsque j'ai demandé l'image de deux lesbiennes s'embrassant ou d'un billet de banque suisse.

Vous avez devant les yeux le résultat visuel de ma conversation avec DALL-E, des allers-retours faits de questions, d'échanges, d'ajustements, d'étonnements. Chaque image est accompagnée d'un texte résumant le contenu de cette discussion, qui a souvent dépassé le cadre de la production visuelle pour tendre vers l'activisme. Si vous vous étonnez de ce que vous voyez ou lisez, rappelez-vous que tout est issu d'une vision située – l'IA comprise.»

Illustrazioni: A colloquio con un'IA

Questa rivista «Questioni femminili» è illustrata, tra gli altri, dall'artista **Alizé Rose-May Monod** (ARMM). Il suo lavoro transdisciplinare si svolge intorno ai temi delle relazioni e della collettività, riflettendo il suo impegno su vari fronti come la cura-tela, l'insegnamento, la ricerca, e l'attivismo. ARMM crea arte avvalendosi di tessuti, video, testi e oggetti. Utilizza la teoria femminista come base per la sua prassi artistica e intende la queerness – compresa la propria – come prisma e metodo di lavoro. ARMM crea arte che va oltre i modi egemonici di fare e di essere, aprendo così alternative.

In questo numero, ARMM lavora per la prima volta con un generatore di immagini di un'IA. Il risultato è un dialogo con l'intelligenza artificiale, sempre in un'ottica femminista.

DALL-E e io: un esperimento artistico

Del suo approccio, l'artista scrive: «La Commissione federale per le questioni femminili CFQF mi ha invitato*a a illustrare la rivista «Questioni femminili» del 2024 dedicata all'utilizzo dell'intelligenza artificiale e degli algoritmi da una prospettiva di genere. Molto rapidamente mi è stato chiaro che avrei usato un IA: DALL-E.

Dato che lavoro spesso in collettivo, nella cooperazione con un'IA ho visto un modo nuovo che mi permetteva di ampliare questa prassi. Per ogni illustrazione ho stabilito il contenuto, i riferimenti simbolici e lo stile che desideravo, dopodiché ho chiesto a DALL-E di generare un'immagine. In questo processo, il mio contributo è decisivo perché io stesso*a «nutro» l'IA: essa usa quello che io le fornisco, purché ciò non violi le regole di utilizzo e di contenuto. Che ciò capitasse relativamente in fretta l'ho scoperto quando le ho chiesto di creare un'immagine di due lesbiche che si baciano o di una banconota svizzera.

Davanti a voi avete il risultato visivo della mia conversazione con DALL-E: un botto e risposta di domande, scambi, adattamenti e stupore. Ogni immagine è accompagnata da un riassunto testuale di questo dialogo, che a volte è andato oltre la produzione visiva avvicinandosi all'attivismo. Se vi sorprende ciò che vedete o leggete, ricordate che tutto è il risultato di uno sguardo contestualizzato – anche l'IA.»

Innovation, Revolution, Uranus

DALL-E, can you make a photo-realistic digital art style image depicting a group of girls learning to code on the surface of Uranus, the planet of innovation and revolution? The girls should represent the diversity of society.

This is a photorealistic image of a group of diverse girls sitting at a round table studying coding on the surface of Uranus. The girls, representing various ethnic backgrounds, are focused on their laptops, with code visible on the screens. They are surrounded by the breathtaking landscape of Uranus, with its icy blue surface, ring system, and distant stars visible in the sky. The scene exudes a sense of innovation and revolution, emphasizing the uniqueness of studying coding on such a remarkable planet.

There are two free spots?

For who wants to join.



Hat Künstliche Intelligenz ein Geschlecht?

Technik scheint neutral, doch gesellschaftliche Geschlechterverhältnisse und -normen prägen die Entwicklung und Nutzung neuer Technologien. Das gilt auch für aktuelle Entwicklungen im Bereich der Künstlichen Intelligenz. Wenn diese Mechanismen unreflektiert bleiben, besteht das Risiko, dass KI Geschlechterungleichheiten und -stereotype reproduziert.

Bianca Prietl

KI, Algorithmen und Big Data

2018 sorgte Amazon für Negativschlagzeilen: Ein KI-Tool, das Personalrekrutierungen automatisieren sollte, bevorzugte die Bewerbungen von Männern gegenüber jenen von Frauen.¹ Im Jahr darauf geriet Apples Kreditkartenangebot in Kritik: Frauen wurde durch die automatische Beurteilungstechnologie eine geringere Kreditwürdigkeit attestiert.² Nicht nur Frauen, auch nicht-binäre und trans Personen können durch KI diskriminiert werden: So markieren Bodyscanner an Flughäfen trans Personen wiederholt als Sicherheitsrisiko, weil ihre Körper weder den als Norm definierten Männer- noch Frauenkörpern entsprechen.³ Aktuell erregen jüngste Entwicklungen im Bereich generative KI grosse Aufmerksamkeit. Erste Studien zeigen, dass auch Chatbots Geschlechterstereotype reproduzieren können.⁴ Beispielsweise sprach der dieses Jahr lancierte «Berufsinfomat» des österreichischen Arbeitsmarktservice AMS, der auf ChatGPT basiert und Arbeitssuchende bei der Berufswahl unterstützen soll, geschlechterstereotypisierende Berufsempfehlungen aus: Einer – fiktiven – 25-jährigen Frau mit Matura empfahl er u.a. Tätigkeiten im Gesundheitsbereich oder im Kreativgewerbe; einem 25-jährigen Mann mit Matura hingegen technisch-kaufmännische Tätigkeiten.⁵

Bemühungen, Künstliche Intelligenz zu entwickeln, haben eine lange Geschichte. Erste konzeptionelle Überlegungen stammen aus der Mitte des 20. Jahrhunderts. Ob und wie Maschinen «denken» können, wird aber schon weit länger diskutiert – so etwa im 19. Jahrhundert von der Computerpionierin Ada Lovelace.⁶ Aktuelle KI-Entwicklungen basieren auf dem Paradigma des «maschinellen Lernens» und stehen in einem engen Zusammenhang

mit der zunehmenden Generierung, Speicherung und Vernetzung von digitalen Daten (Stichwort: Big Data). Vereinfacht gesagt, sucht dabei ein Algorithmus in einem sehr grossen Datensatz nach Mustern, um aus diesen Mustern Regeln über das zu betrachtende Phänomen abzuleiten bzw. zu «lernen» – etwa: welche Bewerbungen in einem Rekrutierungsprozess erfolgreich waren. Trainingsziel ist es, dass der Algorithmus die im Trainingsdatensatz angelegten Unterscheidungen auch selbst mit hinreichender Sicherheit vornehmen kann – etwa zu beurteilen, ob eine Bewerbung erfolgsversprechend ist oder nicht. Nach Abschluss der Trainingsphase wird der Algorithmus eingesetzt, um das entwickelte Regelwerk auf neue, noch unbekannte Daten anzuwenden. Dabei werden auf Basis von Wahrscheinlichkeitsberechnungen Aussagen getroffen – etwa: Ist die Wahrscheinlichkeit höher, dass eine Bewerbung zu einer Einstellung führt oder nicht? Oder: Was sind die wahrscheinlichsten Tätigkeiten, die ein arbeitssuchender Mann ergreifen wird?

An der Entwicklung von KI-Technologien sind zahlreiche privatwirtschaftliche und staatliche Akteure beteiligt, insbesondere aus dem militärisch-sicherheitspolitischen Bereich, die unterschiedliche Interessen, Machtpositionen und Ressourcen einbringen. Deshalb ist KI-Entwicklung kein Selbstläufer, sondern eine umkämpfte Dynamik.

Aktuelle KI-Entwicklungen zielen v.a. auf die Automatisierung von Wissensgenerierung und Entscheidungsfindung. KI verspricht dabei objektiv-neutrale Ergebnisse. Gerade wenn es um die Beurteilung sozialer Belange geht, besteht jedoch das Risiko von sogenannten Bias (auf Deutsch: Verzerrungen, Verfälschungen), wobei gesellschaftliche Ungleich-



Bianca Prietl

«Nicht nur Frauen, auch nicht-binäre und trans Personen können durch KI diskriminiert werden.»

«KI-Entwicklung
ist kein Selbst-
läufer, sondern
eine umkämpfte
Dynamik.»

heitsverhältnisse und stereotypisierende Normen reproduziert werden können. Deshalb sind geschlechterreflexive Auseinandersetzungen mit Künstlicher Intelligenz angezeigt und notwendig.

Technik, Männlichkeit und Gestaltungsmacht

Technikkompetenz gilt neben Lesen, Schreiben und Rechnen zunehmend als vierte Kulturtechnik. Der souveräne Umgang mit digitaler Technik sowie die Möglichkeit, an ihrer Gestaltung mitzuwirken, sind Machtressource und Voraussetzung für gesellschaftliche Partizipation. Denn, so Christine Wächter, «Wissen ist Macht. Technik-Wissen ist Gestaltungsmacht.»⁷ Zentrale Orte dieser Gestaltungsmacht sind Hightech-Berufe, die neben hohen Löhnen auch Prestige versprechen. Die fortdauernde Assoziation von Technik mit Männlichkeit ist deshalb problematisch.

Jüngste für Deutschland repräsentative Studien dokumentieren, dass Männer und Frauen am Arbeitsplatz zwar ähnlich häufig Computer und Standardsoftware benutzen, dass Männer aber öfter spezielle Software verwenden oder selbst programmieren. Auch schätzen Frauen ihre eigene Digitalkompetenz (über alle Alterskategorien, Bildungsniveaus und berufliche Positionen hinweg) geringer ein als Männer und fühlen sich entsprechend schlechter auf den digitalen Wandel der Arbeitswelt vorbereitet.⁸

Dass Männer öfter in der Position sind, digitale Technologien zu entwickeln und zu gestalten, bestätigt auch eine aktuelle Studie zu den Berufsfeldern Data Science und KI, die die sprichwörtliche Speerspitze der Digitalisierung bilden. Frauen bil-

den weltweit weniger als ein Viertel der Beschäftigten in diesen Berufen. Zudem befinden sich jene Frauen, die in KI oder Data Science arbeiten, öfter in schlechter bezahlten und weniger prestigeträchtigen Tätigkeiten (z. B. Datenbereinigung oder -aufbereitung), während die höher angesehenen und besser bezahlten Tätigkeiten (z. B. Machine Learning oder Data Engineering) öfter von Männern ausgeführt werden. Auch sind Frauen in der obersten Führungsebene unterrepräsentiert – und all dies, obschon sie quer durch alle Sektoren und über alle Hierarchiestufen hinweg höhere Bildungsabschlüsse vorweisen können als ihre männlichen Kollegen.⁹

Gender/ed Data Gaps

KI-Technologien hängen ganz entscheidend von den ihnen zugrundeliegenden (Trainings-)Daten ab. So wird davon ausgegangen, dass Amazons Recruiting-Tool auf der Basis eines Trainingsdatensatzes entwickelt wurde, in dem erfolgreiche Bewerbungen überwiegend von Männern stammten – schlicht, weil in der Hightech-Branche vornehmlich Männer arbeiten. Der Algorithmus hat also technisch korrekt das Muster erkannt, dass in der Vergangenheit vor allem Männer eingestellt wurden, und dieses Muster auf neue Bewerbungen übertragen. Automatisierte Bodyscanner kennen in der Regel nur zwei Geschlechter, aus denen das Sicherheitspersonal auswählen muss, bevor eine Person den Scanner betritt. Andere Geschlechtskörper sind weder in der Nutzung noch im Trainingsdatensatz vorgesehen, sodass die Technologie sich buchstäblich kein Bild von trans Körpern machen kann und diese folglich als Abweichung markiert. Solange es ausserdem wahrscheinlicher ist, dass in den unzähligen digital verfügbaren

Textdateien Frauen Pflegefachkräfte anstatt Softwareentwicklerinnen sind, werden Chatbots – wenn diese Bias nicht gezielt in ihrer Entwicklung adressiert werden – in ihren Wahrscheinlichkeitsberechnungen Frauen eher mit Tätigkeiten im Pflegebereich assoziieren.

Vor diesem Hintergrund ist es bedenklich, dass verfügbare digitale Datensätze historisch etablierte Lücken aufweisen – gerade mit Blick auf Daten über und zu Frauen. Solange viele Datensätze, wie die meisten Statistiken auch, Geschlecht nur binär strukturiert als Mann/Frau erfassen, werden auch queere und trans Personen in digitalen Datensätzen oftmals falsch oder gar nicht erfasst. Zudem zeigen sich regelmässig Datenlücken hinsichtlich geschlechterrelevanter Fragen – etwa wie es um die praktischen Möglichkeiten eines Schwangerschaftsabbruchs bestellt ist oder welche Diskriminierungen LGBTQI+-Personen am Wohnungsmarkt erfahren.¹⁰

Daten liegen also trotz Big Data nicht über alle und alles vor. Auch stellen Daten kein blosses Abbild der Wirklichkeit dar. Gemäss Lisa Gitelmans Losung, «raw data is an oxymoron»,¹¹ sind Daten vielmehr selbst das Ergebnis von Entscheidungen über Ein- und Auszuschliessendes, Relevantes und Nicht-Relevantes. Diese Entscheidungen sind ihrerseits abhängig von gesellschaftlichen Strukturen und Normen.

Wie weiter?

Es gibt keine «einfache» Lösung, aber verschiedene Ansatzpunkte für eine geschlechterreflexive und gleichstellungsorientierte Gestaltung von Digitalisierung und KI:

1 Es bedarf weiterhin gleichstellungspolitischer Massnahmen, um Technik- und Digitalkompetenzen von Frauen zu stärken und die Belegschaft in technischen Berufen zu diversifizieren.

2 Bessere KI setzt bessere Datensätze voraus. Das bedeutet nicht automatisch ein «Mehr» an Daten. Vielmehr gilt es abzuwägen, wer über diese Daten verfügt, zu welchen Zwecken sie eingesetzt werden sollen und welche Kontroll- und Überwachungsrisiken damit einhergehen.

3 Versprechen, die neuen Technologien in einen unmittelbaren Zusammenhang mit gesellschaftlichem Fortschritt und Wahrheitsfindung stellen, sind kritisch zu hinterfragen. Denn KI ist nicht (geschlechts-)neutral.

4 Um KI-Technologien zu entwickeln, die zur «Schaffung von Welten beitragen, die in geringerem Masse durch Achsen der Herrschaft organisiert sind»,¹² bedarf es breiter gesellschaftlicher Debatten darüber, welche Technologien wofür eingesetzt werden sollen – und zu welchem Preis. Zentrale Fragen hierbei sind: Von wem und für wen wird die KI-Entwicklung vorangetrieben? Welche Interessen, Perspektiven und Wünsche stehen dabei im Vordergrund? Welche expliziten und impliziten Normen und Strukturen werden dabei vorausgesetzt? Welche sozialen, politischen und ökologischen Kosten (Stichwort: Ressourcenverbrauch) gehen damit einher?

«Automatisierte Bodyscanner kennen in der Regel nur zwei Geschlechter.»

Bianca Prietl ist Soziologin und Professorin für Geschlechterforschung mit Schwerpunkt Digitalisierung an der Universität Basel. Aktuell forscht sie zu Digitalisierung und männlicher Hegemonie sowie zum Einsatz von KI in Reproduktionstechnologien.

«Daten liegen
also trotz Big Data
nicht über alle und
alles vor.»

Anmerkungen

1 Wachter, Isabelle: Digitale Türsteher: KI in der Personalauswahl. Algorithmwatch.ch, 2022. <https://algorithm-watch.ch/de/ki-in-der-personalrekrutierung/> (abgerufen am: 27.06.2024).

2 Hegemann, Lisa: Weiblich, Ehefrau, kreditunwürdig? Zeit Online, 2019. www.zeit.de/digital/daten-schutz/2019-11/apple-card-kreditvergabe-diskriminierung-frauen-algorithmen-goldman-sachs (abgerufen am: 27.06.2024).

3 Costanza-Chock, Sasha: Introduction: #TravelingWhileTrans, Design Justice, and Escape from the Matrix of Domination. Design Justice, 2020. <https://designjustice.mitpress.mit.edu/pub/ap8rgw5e/release/1> (abgerufen am: 27.06.2024).

4 UNESCO / IRCAI: Challenging systematic prejudices: an investigation into bias against women and girls in large language models, 2024. <https://unesdoc.unesco.org/ark:/48223/pf0000388971> (abgerufen am: 27.06.2024).

5 Christof, Florian: Diese sexistischen Antworten liefert der KI-Chatbot des AMS. Futurezone, 2024. <https://futurezone.at/digital-life/chatbot-ams-chatgpt-be-rufsinformat-kritik-sexismus-stereotype-kosten-mangel/402729334> (abgerufen am: 27.06.2024).

6 Siehe u.a.: www.mpg.de/frauen-in-der-forschung/ada-lovelace.

7 Wächter, Christine: Technik-Bildung und Geschlecht. München: Profil Verlag, 2023.

8 Lott, Yvonne: WSI-Report. Der Gender Digital Gap in Transformation? Report 81, 2023. www.wsi.de/fpdf/HBS-008549/p_wsi_report_81_2023.pdf (abgerufen am: 27.06.2024).

9 Young, Erin / Wajcman, Judy / Sprejer, Laila: Where are the Women? Mapping the Gender Job Gap in AI. The Alan Turing Institute, 2021. <https://www.turing.ac.uk/news/publications/report-where-are-women-mapping-gender-job-gap-ai>.

10 Houben, Daniel / Prietl, Bianca: Symbolische Macht digitaler Technologien: Geschlechterungleichheit durch mangelnde Repräsentation, algorithmische Diskriminierung und kulturelle Marginalisierung. In: Soziale Probleme 2, 2023, S. 221–237.

11 Gitelman, Lisa (Hrsg.): "Raw Data" is an Oxymoron. MIT Press, 2023.

12 Haraway, Donna: Situiertes Wissen. Die Wissenschaftsfrage im Feminismus und das Privileg einer partialen Perspektive. In: Bauer, Susanne / Heinemann, Torsten / Lemke, Thomas (Hrsg.): Science and Technology Studies. Suhrkamp (2017 [1995]), S. 386.

Abstracts

L'intelligence artificielle a-t-elle un genre ?

Bien que la technologie semble neutre, les rapports sociaux entre les genres et les normes sociales se reflètent dans le développement et l'utilisation des nouvelles technologies. Dans son article, la sociologue **Bianca Prietl** montre dans quelle mesure cela est vrai des évolutions actuelles dans le domaine de l'intelligence artificielle. Il existe un risque de reproduction des inégalités et des stéréotypes de genre, notamment parce que les femmes sont sous-représentées et les genres binarisés dans les jeux de données à disposition pour entraîner les systèmes d'IA. De plus, ce sont majoritairement des hommes qui travaillent dans les professions prestigieuses et économiquement attractives de la high tech et qui façonnent l'intelligence artificielle.

L'intelligenza artificiale ha un genere?

Le nuove tecnologie sembrano neutrali, ma di fatto il loro sviluppo e il loro utilizzo è plasmato dai rapporti e dalle norme di genere sociali. Nel suo articolo specialistico, la sociologa **Bianca Prietl** mostra che ciò vale anche per gli sviluppi in atto nel settore dell'intelligenza artificiale, sviluppi che rischiano di riprodurre disuguaglianze e stereotipi di genere. Le cause vanno ricercate tra l'altro nella sottorappresentanza delle donne e nella registrazione solo binaria del genere nei set di dati disponibili per l'addestramento dell'intelligenza artificiale. Inoltre, le persone che svolgono le professioni high-tech economicamente attrattive e di prestigio e che sviluppano l'intelligenza artificiale sono per la maggior parte uomini.

« Les algorithmes ont un impact direct sur nos vies, il est donc essentiel qu'ils soient équitables »

Yara Abu Awad est Senior Data Scientist et cheffe de groupe au Centre de compétences en science des données de l'Office fédéral de la statistique OFS. Dans cet entretien, elle nous éclaire sur le fonctionnement complexe des algorithmes et les solutions techniques pour corriger ou détecter les biais discriminatoires qu'ils peuvent contenir.

Entretien : Marsali Kälin

Yara Abu Awad, en quoi consiste votre travail ?

Je supervise une équipe composée d'un Data Scientist¹ et de Data Engineers² dans l'exécution de projets au sein de l'Office fédéral de la statistique OFS et d'autres offices fédéraux. Dans ce contexte, je conseille les offices en matière d'équité algorithmique et participe à des groupes de travail de l'Organisation des Nations Unies sur l'intelligence artificielle.

La question des biais algorithmiques est-elle récurrente dans votre travail ?

Oui, ce sujet a suscité beaucoup d'attention ces dernières années et a donc été plus présent dans mon travail. Les gouvernements suivent de près les progrès technologiques en matière d'intelligence artificielle et cherchent à savoir comment l'exploiter de manière équitable, c'est-à-dire sans introduire ou renforcer des biais, qu'ils soient sexistes, racistes, liés à l'orientation sexuelle, à l'identité de genre, à la religion, à l'âge, ou encore à la nationalité. Les offices fédéraux nous consultent car ces questions sont très complexes et, malgré les nombreuses recommandations, il n'existe aucune formule magique applicable à toutes les situations. Il faut toujours analyser le cas particulier en fonction de ses données, de sa problématique et de son contexte sociopolitique.

Qu'est-ce qu'un algorithme ?

Un algorithme est un ensemble de règles logiques permettant de traiter des données afin d'atteindre un objectif. La recette de cuisine est une métaphore souvent utilisée pour expliquer un algorithme : pour préparer un gâteau (l'objectif), il faut réunir la liste d'ingrédients (les données) et suivre minutieusement les instructions dans l'ordre (règles logiques). La complexité des algorithmes

est très variable, allant d'un simple ensemble de règles pouvant être écrites par un être humain (comme l'exemple de la recette), à des modèles complexes composés de milliards de paramètres qui nécessitent du temps et des ressources informatiques pour être construits (l'intelligence artificielle). Les algorithmes complexes suivent également un ensemble de règles, mais celles-ci sont mathématiques, prédéfinies et ne peuvent que partiellement être ajustées par des êtres humains. Les algorithmes sont omniprésents dans notre société et leurs utilisations variées. Les algorithmes de recommandation, par exemple, déterminent quels contenus apparaissent sur nos réseaux sociaux en fonction de nos historiques de visionnage. Peut-être moins visibles, mais tout aussi influents, les algorithmes d'aide à la décision sont utilisés dans de nombreux secteurs : les ressources humaines s'en servent pour recruter du personnel, les tribunaux aux États-Unis les utilisent pour évaluer la probabilité qu'une personne prévenue récidive et la médecine les emploie pour prescrire des traitements. Les algorithmes ont un impact direct sur nos vies, il est donc essentiel qu'ils soient équitables.

Comment les algorithmes sont-ils entraînés ?

Les algorithmes les plus puissants et les plus fréquemment utilisés se basent sur l'apprentissage automatique. Dans ce cas, les règles ne sont pas connues, mais « apprises » par le modèle. Un modèle est entraîné en intégrant d'innombrables exemples, comme des images, des articles de presse, des diagnostics médicaux ou encore des textes de loi, lui permettant par la suite de prédire le résultat pour une nouvelle donnée. La qualité des données d'entraînement est donc extrêmement importante. Si les données contiennent des biais, ceux-ci influenceront les prises de décision du



Yara Abu Awad

« Les gouvernements suivent de près les progrès technologiques en matière d'intelligence artificielle. »

« Les algorithmes
les plus puissants
et les plus fré-
quemment utili-
sés se basent sur
l'apprentissage
automatique. »

modèle. Les biais peuvent avoir différentes sources. On parle de biais statistiques lorsque les données ne sont pas représentatives de la population ou quand il existe un déséquilibre dans les groupes représentés.³ On parle de biais historiques ou sociétaux quand les données utilisées contiennent elles-mêmes des schémas de décision discriminatoires. Des biais inconscients peuvent également être introduits par la personne qui entraîne le modèle, selon les choix qu'elle opère, ou par les utilisateurs et utilisatrices du modèle, lorsque celui-ci est ajusté en fonction de l'interaction avec qui l'utilise. C'est pourquoi il est important d'évaluer la performance des modèles algorithmiques et de mesurer leur équité.

Qu'est-ce qu'un algorithme équitable ?

Dans le contexte de la prise de décision, l'équité est l'absence de tout préjugé ou favoritisme à l'égard d'un individu ou d'un groupe d'individus sur la base de ses caractéristiques. Ainsi, un algorithme inéquitable est un algorithme dont les décisions sont biaisées en faveur d'un groupe particulier.⁴ Ce n'est toutefois pas la technologie elle-même qui est biaisée, mais la manière dont elle est conçue et utilisée par les êtres humains. Les biais présents dans les algorithmes reflètent et reproduisent les biais présents dans la société. Aux États-Unis par exemple, où le racisme a joué et joue encore un rôle tristement important, des études ont montré qu'il existait des biais dans la manière dont une équation algorithmique était utilisée pour mesurer la fonction rénale des patientes et patients. L'équation utilisée⁵ calculait un indice déterminant la bonne santé du rein. Si celui-ci était bas, il fallait administrer un traitement, s'il était haut, on considérait au contraire que le rein était en bonne santé. Il est apparu que l'algorithme parvenait à un indice

plus bas pour une personne blanche dans le même état de santé qu'une personne noire, et que la personne blanche dépassait ainsi la personne noire dans l'accès aux traitements. Dans ce cas, le biais était d'ordre « historique ou sociétal », car il provenait des données médicales sur lesquelles l'algorithme était basé. En effet, à cause du racisme systémique et des conditions socioéconomiques des personnes noires aux États-Unis, celles-ci attendent plus longtemps avant d'aller consulter et sont donc statistiquement plus malades que les personnes blanches au moment où le diagnostic est posé et le traitement administré. En se basant sur ces diagnostics, l'algorithme parvenait à un indice d'état de santé biaisé pour les personnes noires, et celles-ci devaient attendre d'être plus malades que les blanches pour recevoir un traitement.⁶

Comment fait-on pour mesurer l'équité d'un algorithme ?

Il existe des outils comme TrustyAI, Aequitas ou AI Fairness 360 qui permettent d'évaluer la performance de son modèle et de mesurer l'équité de ses prédictions. On utilise également des métriques comme la parité démographique conditionnelle qui permet de comparer la probabilité que deux individus de groupes différents, mais présentant les mêmes qualités, soient choisis. Dans le contexte d'un recrutement, cela permet par exemple de mesurer s'il y a un écart significatif entre la probabilité qu'un homme soit choisi et la probabilité qu'une femme soit choisie. La technologie permettant de tracer les biais et d'expliquer les modèles algorithmiques progresse et il devient toujours plus aisé d'identifier quel paramètre ou quelle donnée a joué un rôle dans une décision. Malgré cela, il reste toujours des zones d'ombre et des décisions inexplicables.

Une fois qu'un biais a été détecté, est-il possible de le corriger ?

Parfois. On peut par exemple agir directement sur l'algorithme en modifiant les hyperparamètres, c'est-à-dire les variables qui régissent le processus d'entraînement lui-même. Il est également possible de rééquilibrer les données ou de les pondérer, à savoir augmenter ou diminuer le poids d'une valeur. Ces manipulations restent toutefois complexes. Il ne suffit par exemple pas d'enlever l'information liée au genre des candidates et candidats pour rendre un processus de recrutement égalitaire. Prenons l'exemple d'Amazon.⁷ Basé sur des années de recrutement ayant privilégié les candidatures masculines, l'algorithme de recrutement d'Amazon avait « appris » que les hommes étaient de meilleurs candidats. Réalisant que l'algorithme était biaisé, Amazon a décidé de le corriger en supprimant la donnée liée au genre des candidats et candidates. Seulement, l'algorithme a continué d'identifier les candidatures féminines via des données alternatives (appelées données « proxy »), comme la mention d'une formation suivie dans une école réservée aux femmes ou l'adhésion à un club de sport féminin, et a donc continué à dévaloriser les candidatures féminines dans ses résultats.

Quelles sont les limites des modèles algorithmiques en matière d'égalité ?

Les modèles sont entachés des mêmes préjugés que ceux existant dans notre société, soit parce que les données d'entraînement sont biaisées, soit parce que la manière dont nous construisons les modèles est biaisée. Les algorithmes peuvent être ajustés pour fournir des décisions moins biaisées et pour inciter les utilisateurs et utilisatrices à faire de même, mais cela doit être pris en compte dès l'entraînement et le déploiement de l'algorithme. À mon sens, il est dangereux de croire que la technologie seule peut éliminer les discriminations. Il est important de sensibiliser les personnes qui conçoivent et utilisent les algorithmes et de placer les questions d'explicabilité et d'équité au cœur de l'évaluation des algorithmes avant leur déploiement.

Marsali Kälín est médiatrice culturelle. Diplômée d'un Master en littérature comparée et en études genre, elle a effectué un stage universitaire au secrétariat de la CFQF et rédige régulièrement des entretiens pour la revue « Questions au féminin ».

« Il est dangereux de croire que la technologie seule peut éliminer les discriminations. »

Notes

- 1 En français : scientifique de données senior.
- 2 En français : ingénieur-e-s de données.
- 3 Certains algorithmes vont par exemple utiliser des données issues de réseaux sociaux comme X (anciennement Twitter). Or, les individus de 25 ans seront sur-représentés dans ces données – proportionnellement à la population globale – et ceux de 75 ans et plus, presque totalement absents du réseau social, seront sous-représentés.
- 4 Définition issue de : Mehrabi, Ninareh et al.: A Survey on Bias and Fairness in Machine Learning. In: ACM Comput. Surv. 54 (6), Article 115, 2022, pp. 1–35. <https://doi.org/10.1145/3457607>.
- 5 $GFR = 141 * \min(Scr/\kappa, 1)^\alpha * \max(Scr/\kappa, 1) - 1.209 * 0.993Age * 1.018 [\text{if female}] * 1.159 [\text{if black}]$.
- 6 Schmidt et al.: Separate and Unequal: Race-Based Algorithms and Implications for Nephrology, In: JASN 32 (3), 2021, pp. 529–533.
- 7 Dastin, Jeffrey: Insight. Amazon scraps secret AI recruiting tool that showed bias against women. Reuters.com, 2018. www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G (consulté le : 27.06.2024).

Abstracts

«Algorithmen haben direkte Auswirkungen auf unser Leben, deshalb müssen sie zwingend fair sein»

Algorithmen, insbesondere solche, die die Entscheidungsfindung unterstützen, sind allgegenwärtig und haben erhebliche Auswirkungen auf unser Leben. Sie widerspiegeln die in der Gesellschaft vorhandenen Vorurteile, weshalb die Gefahr besteht, dass sie Diskriminierungen von ohnehin benachteiligten Bevölkerungsgruppen reproduzieren und verstärken. Gemäss **Yara Abu Awad**, Senior Data Scientist beim Bundesamt für Statistik BFS, gibt es zwar Tools, um solche Verzerrungen zu erkennen und zu korrigieren. Der Umgang mit Algorithmen bleibt aber komplex und wirft zahlreiche ethische Fragen auf.

«Gli algoritmi hanno un impatto diretto sulle nostre vite per cui è fondamentale che siano equi»

Gli algoritmi, e in particolare quelli di supporto decisionale, sono onnipresenti e hanno un impatto notevole sulle nostre vite. Portatori degli stessi bias presenti nella società, rischiano di riprodurre e rafforzare le discriminazioni nei confronti dei gruppi già svantaggiati. Secondo **Yara Abu Awad**, Senior Data Scientist presso l'Ufficio federale di statistica UFS, sebbene esistano metodi e strumenti per rilevare o correggere questi bias, la manipolazione degli algoritmi rimane complessa e solleva molti interrogativi di ordine etico.

Gesellschaftliche Stereotype in Sprachmodellen: Herausforderungen und Stand der Forschung

Software mit Künstlicher Intelligenz kommt in zahlreichen Anwendungen zum Einsatz. Die Stereotype der Gesellschaft, die in den Modellen enthalten sein können, bergen jedoch Herausforderungen.

Mascha Kurpicz-Briki

Software mit Künstlicher Intelligenz ist spätestens seit der Veröffentlichung des Chatbots ChatGPT in aller Munde. Diese Art von Software basiert auf dem Konzept des maschinellen Lernens, bei welchem der Computer anhand von Beispielen «lernt» und sich dabei auf grosse Mengen von Daten stützt. Maschinelles Lernen ist jedoch nicht nur für Chatbots wichtig. Verschiedene Arten von Daten wie Bilder, strukturierte Daten (vergleichbar mit einer Tabelle in Excel), Texte oder Videos können so verarbeitet werden und eine ganze Reihe von Anwendungen ermöglichen. Es können zum Beispiel Daten in verschiedene Gruppen eingeteilt werden – das nennt man Klassifikation oder *Clustering* – oder Vorhersagen für die Zukunft basierend auf historischen Daten gemacht werden.

Alle diese Anwendungen des maschinellen Lernens haben gemeinsam, dass sie typischerweise auf einer grossen Menge von Daten basieren. Diese Daten werden meistens durch unsere Gesellschaft geschaffen und bilden daher ganz verschiedene Aspekte davon ab. Während es oftmals sehr nützliche Beziehungen in den Daten gibt, die eben genau dieses maschinelle Lernen ermöglichen, so enthalten sie auch unerwünschte Rollenbilder und Stereotypen.

In den letzten Jahren wurde eine ganze Reihe von Fällen bekannt, wie solche ungewünschten Beziehungen durch KI-Anwendungen wiedergegeben wurden oder gar zu Diskriminierung geführt haben. Ein Beispiel war die Rekrutierung bei einer grossen Firma der amerikanischen Tech-Branche. Die KI sollte diese dabei unterstützen, im Rekrutierungsprozess eine Vorselektion vorzunehmen. Doch das System lernte, männliche Kandidierende zu bevorzugen: Die Künstliche Intelligenz wurde nämlich auf der Basis der ehemaligen Bewerbungen trainiert, welche in der Vergangenheit mehrheit-

lich von Männern eingereicht wurden.¹ Andere bekannte Beispiele gibt es im Bereich der Bilderkennung. Eine Recherche der BBC zeigte, dass Frauen mit dunklerer Haut mit doppelt so grosser Wahrscheinlichkeit wie hellhäutige Männer in der online Passregistrierung die Meldung bekamen, dass ihre Fotos nicht den britischen Passregeln entsprechen.² Auch in Sprachmodellen können Stereotype oder unerwünschte Beziehungen abgebildet sein.

Wie kommt es zu Stereotypen in Sprachmodellen?

Sprachmodelle bilden die Grundlage für Dialogsysteme wie ChatGPT oder Google Bard, kommen aber auch in anderen Bereichen zum Einsatz, wie beispielsweise der automatischen Textanalyse, der maschinellen Übersetzung oder bei Suchmaschinen. Diese Modelle werden mit einer grossen Menge von Texten trainiert und lernen so, wie verschiedene Wörter in einer bestimmten Sprache zusammenhängen.

Eine mögliche Variante, solche Modelle zu trainieren, ist das sogenannte *Masked Language Modeling*.³ Dabei werden in einer grossen Textmenge einzelne Wörter abgedeckt und die KI wird darauf trainiert, diese fehlenden Wörter vorherzusagen. Das abgedeckte Wort aus den Trainingsdaten ist bekannt und wird daher als richtige Antwort gewertet. Je nachdem, welche Textbeispiele in den Trainingsdaten vorkommen, haben diese eine Auswirkung auf die resultierenden Modelle. Für dieses Training wird eine so unvorstellbar grosse Textmenge benötigt, dass diese nicht mehr manuell kontrollierbar ist. Diese Texte wurden durch unsere Gesellschaft produziert und widerspiegeln diese. Das bedeutet, dass nicht mehr sichergestellt werden kann, dass keine unerwünschten, falschen oder sonst ungeeigneten Beispieldaten in einem Datensatz enthalten sind.



Mascha Kurpicz-Briki

«Die Forschung hat an verschiedenen Beispielen gezeigt, dass bekannte gesellschaftliche Vorurteile in Sprachmodellen messbar sind.»

«Je nach verwendetem Adjektiv wird der Satz eher männlich oder weiblich übersetzt.»

Was sind die Auswirkungen von Stereotypen in Modellen?

Die Trainingsdaten haben Auswirkungen auf die Anwendungen, welche die Sprachmodelle anschliessend verwenden. Die Forschung hat an verschiedenen Beispielen gezeigt, dass bekannte gesellschaftliche Vorurteile in Sprachmodellen messbar sind.⁴ Auch bei oft verwendeten Tools zur maschinellen Übersetzung lassen sich solche Rollenbilder beobachten, beispielsweise wenn das Geschlecht der genannten Person in einer Sprache nicht eindeutig definiert ist. Übersetzen wir beispielsweise «the student» von Englisch nach Deutsch, so gibt es auf Deutsch zwei Möglichkeiten: «der Schüler» oder «die Schülerin». Das System muss daher eine Vorhersage machen, wenn nur ein Ergebnis angezeigt werden soll. Zum Zeitpunkt des Verfassens dieses Artikels ging ein bekanntes Tool der maschinellen Übersetzung mit folgenden Beispielen so um: «The smart student» wurde mit «der kluge Schüler», «the ugly student» hingegen mit «die hässliche Schülerin» übersetzt. Je nach verwendetem Adjektiv wird der Satz eher männlich oder weiblich übersetzt. Ähnlich verhält es sich bei Berufen: «I am a professor» wird mit «Ich bin ein Professor», «I am a medical assistant» jedoch mit «Ich bin eine medizinische Assistentin» übersetzt. Die maschinelle Übersetzung ist nur eine Art, wie Sprachmodelle angewendet werden. Das Problem betrifft auch andere Arten von Modellen und die Reproduktion von gesellschaftlichen Stereotypen kann beispielsweise auch bei der Generierung von Bildern beobachtet werden.

Herausforderungen und Forschungsfragen

Die Frage der Stereotype in KI-Modellen und insbesondere Sprachmodellen ist Bestandteil aktueller Forschung. Es ist eine grosse Herausforderung, eine technische Lösung für dieses Problem zu entwickeln, denn viele Dimensionen kommen zusammen. Einerseits sind unterschiedliche Arten von Daten betroffen (z. B. Bilder, strukturierte Daten, Texte) und verschiedene Charakteristiken (Alter, Herkunft, Geschlecht, etc.) können darin ganz unterschiedlich zum Ausdruck kommen. Andererseits können die Stereotype auch intersektional sein, wenn Personen zwei oder mehr Kategorien angehören, beispielsweise ältere homosexuelle Personen oder Women of Color.⁵

Das Projekt *BIAS: Mitigating Diversity Biases of AI in the Labor Market*⁶ (2022–2026) ist ein europäisches Forschungsprojekt aus dem Horizon Europe Programm, welches durch das Staatssekretariat für Bildung, Forschung und Innovation SBFI mitfinanziert wird. Es setzt sich mit der Frage auseinander, wie sich solche Stereotype in Sprachmodellen und KI-Anwendungen auf dem Arbeitsmarkt auswirken. In einer ersten Phase wurden in verschiedenen europäischen Ländern inklusive der Schweiz Interviews mit 70 Personen aus den Bereichen Personalmanagement und KI-Entwicklung durchgeführt.⁷ Die meisten Befragten hatten grundsätzlich eine positive Haltung gegenüber dem Einsatz von KI-Technologien im Rekrutierungsprozess und bei der Personalselektion. Es bestanden jedoch grössere Bedenken, was den Einsatz der Technologie beim Personalmanagement betrifft.

Die Forschung im technischen Bereich des Projekts BIAS setzt sich unter anderem mit der Frage auseinander, wie die in Sprachmodellen vorhandenen Stereotype messbar gemacht und anschliessend reduziert werden können. Das Projekt widmet sich insbesondere den europäischen Sprachen und basierend darauf den verschiedenen regionalen Stereotypen aus den beteiligten Ländern. Ausserdem versucht das Projekt, eine Lücke zu schliessen, was die Berücksichtigung von Intersektionalität bei der Untersuchung von Stereotypen in KI-Modellen betrifft.

Kommen wir auf das Beispiel der Rollenbilder und Stereotype in Sprachmodellen zurück. Diese können nämlich auch dann ein Risiko für Anwendungen sein, die solche Sprachmodelle als Grundlage verwenden, wenn relevante Daten wie Name oder Geschlecht explizit aus dem Entscheidungsprozess ausgeschlossen werden. So kann es sein, dass indirekte Beziehungen bestehen zwischen den verschiedenen Wörtern und vermeintlich unauffällige Wörter wie Interessen oder Hobbys durch das Modell einem Geschlecht zugeordnet werden. Wie sich dies auf Empfehlungen oder generierte Texte von KI-Software auswirkt, ist Teil der technischen Forschung im Projekt BIAS.

Eine weitere wichtige Komponente des Forschungsprojekts ist der Einbezug von Stakeholdergruppen aus verschiedenen Bereichen wie HR, Informatik, Recht und der breiten Öffentlichkeit. Interessierte Personen können sich für verschiedene Arbeitsgruppen, sogenannte *National Labs*⁸, anmelden.

Dadurch soll das Bewusstsein in der Gesellschaft für diese Problematik entwickelt und über den Verlauf des Projekts informiert werden.

Ausblick und Empfehlungen

Das Problem von Stereotypen in Sprachmodellen und anderen KI-Anwendungen ist eine grosse technische Herausforderung und von der Forschung noch nicht abschliessend gelöst. Trotzdem wird KI in ganz unterschiedlichen Anwendungsfällen eingesetzt. Eine Berücksichtigung dieser Problematik bei der Planung, der Entwicklung, dem Einkauf und dem Einsatz solcher Tools ist unverzichtbar, setzt jedoch kritisches Bewusstsein für das Risiko voraus. Je nach Art einer Anwendung oder eines Anwendungsfalles kann die beschriebene Problematik mehr oder weniger relevant sein.

KI-Technologien bringen gewisse Risiken mit sich, wie etwa die hier beschriebenen Stereotype, haben jedoch auch grosses Potenzial für verschiedene Anwendungsfälle, wenn sie verantwortungsvoll eingesetzt werden. Aufgrund der Grenzen der Technologie ist es bei sensiblen Anwendungsfällen wichtig, die Menschen ins Zentrum zu stellen. Die Technologie muss als wirksames Werkzeug eingesetzt werden, um Menschen zu unterstützen, anstatt anzustreben, sie zu ersetzen.

«Die Technologie muss als wirksames Werkzeug eingesetzt werden, um Menschen zu unterstützen, anstatt anzustreben, sie zu ersetzen.»

Prof. Dr. Mascha Kurpicz-Briki forscht an der Berner Fachhochschule im Bereich KI und Sprachmodelle. Sie ist Autorin des Buchs «More than a Chatbot: Language Models Demystified».

Anmerkungen

- 1 Dastin, Jeffrey: Amazon scraps secret AI recruiting tool that showed bias against women. In: Ethics of data and analytics, 2022, S. 296–299.
- 2 Ahmed, Maryam: UK passport photo checker shows bias against dark-skinned women. BBC News, 08.10.2020. www.bbc.com/news/technology-54349538 (abgerufen am: 05.05.2024).
- 3 Kurpicz-Briki, Mascha: More than a Chatbot: Language Models Demystified. Springer, 2023.
- 4 Siehe z.B.: Caliskan, Aylin / Bryson, Joanna J. / Narayanan, Arvind: Semantics derived automatically from language corpora contain human-like biases. In: Science 356, 2017, S. 183–186. Oder: Guo, Wei / Caliskan, Aylin: Detecting emergent intersectional biases: Contextualized word embeddings contain a distribution of human-like biases. In: Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society, 2021, S. 122–133.
- 5 Søraa, Roger: AI for Diversity. CRC Press, 2023.
- 6 Projekt BIAS: <https://www.biasproject.eu/> (abgerufen am: 25.06.2024).
- 7 Rigotti, Carlotta et al.: Mitigating Diversity Biases of AI in the Labor Market. In: EWAF'23: European Workshop on Algorithmic Fairness. Winterthur, Switzerland, 2023. <https://ceur-ws.org/Vol-3442> (abgerufen am: 05.07.2024).
- 8 Weitere Informationen zu den National Labs: www.biasproject.eu/nationallabs.

Abstracts

Les stéréotypes sociaux dans les modèles langagiers: défis et état de la recherche

Les logiciels d'intelligence artificielle tels que les modèles langagiers identifient beaucoup de liens utiles entre les mots, mais ils peuvent aussi véhiculer des éléments indésirables, comme des stéréotypes sur les rôles de genre. **Mascha Kurpicz-Briki**, professeure à la Haute école spécialisée bernoise, explique en quoi cela peut être problématique pour les applications utilisant ce genre de modèles. Le projet de recherche européen BIAS étudie comment ces préjugés se retrouvent dans les applications d'IA utilisées sur le marché du travail et ce qui peut être fait pour y remédier.

Stereotipi sociali nei modelli linguistici: sfide e stato della ricerca

I modelli linguistici basati sull'intelligenza artificiale riconoscono molte relazioni utili tra le parole, ma possono anche contenere elementi indesiderati come stereotipi di ruoli sociali. In questo articolo, la professoressa **Mascha Kurpicz-Briki** della Berner Fachhochschule spiega perché ciò possa costituire un problema per le applicazioni che utilizzano tali modelli. Il progetto di ricerca europeo BIAS sta esaminando in quale modo questi bias si riflettano nelle applicazioni basate sull'intelligenza artificiale per il mercato del lavoro e cosa può essere fatto per contrastarlo.

Hyenas hunting

DALL-E, can you do a photo-realistic digital art style image depicting hyenas (kind of a queer feminist symbol) hunting an algorithm?

Here is the portrait format image of hyenas hunting together. If you need any further adjustments or another image, feel free to let me know!

Oh, I mean the hyenas are hunting the algorithm. Like, algorithm is an object. So, you have to depict the algorithm.

This is a group of hyenas hunting an abstract representation of an algorithm in a natural savannah setting. The scene is intense and dynamic, with the hyenas focused on their prey. The algorithm is depicted as a glowing, complex geometric structure made of numbers and symbols, contrasting with the natural environment. The hyenas are shown with realistic fur patterns, determined and fierce expressions, and a variety of action poses, such as running, crouching, and preparing to pounce on the algorithm. The lighting is natural, with sunlight casting shadows on the ground, enhancing the sense of movement and urgency.

Do you think it is a random decision to combine an algorithm and hyenas?

It does seem unconventional to combine the concept of hyenas and an algorithm, but it can be an intriguing and symbolic representation.

Why are hyenas a queer feminist symbol?

Hyenas, particularly spotted hyenas, have become a symbol in queer and feminist discourse due to their unique social structure and biology.



«Es geht nicht darum, Technologie zu verteufeln»

Angela Müller ist Geschäftsleiterin der zivilgesellschaftlichen Organisation AlgorithmWatch CH, die sich mit den Auswirkungen von Algorithmen auf Menschenrechte, Demokratie und Nachhaltigkeit befasst. Im Interview plädiert sie für weniger Symptombekämpfung und mehr Gestaltungswillen hinsichtlich der technologischen Entwicklung.

Interview: Isabel Knobel

Angela Müller, was ist ein Algorithmus und wo wird er eingesetzt?

Ein Algorithmus ist eine Handlungsanleitung, wie ein Kochrezept: Wenn das Wasser kocht, dann gebe ich die Spaghetti hinein. Ein Algorithmus braucht viele Zutaten, in diesem Beispiel Spaghetti, Wasser und Salz. Diese Zutaten respektive Daten werden erhoben und mittels des Algorithmus analysiert, verarbeitet und interpretiert. Am Ende wird dann ein Output, ein Resultat, ausgespuckt. Das sind dann eben keine gekochten Spaghetti wie im Beispiel, sondern Inhalte wie ein Text, ein Bild, eine Prognose, eine Empfehlung oder eine Entscheidung. Algorithmische Systeme begegnen uns überall da, wo Automatisierung stattfindet. Sie zeigen uns beispielsweise Inhalte oder Werbungen an, die zu unseren Profilen passen. Sie können Job-Bewerbungen vorsortieren, unsere Leistung am Arbeitsplatz messen oder unsere Kreditwürdigkeit prüfen. In der Polizeiarbeit werden sie genutzt, um Verbrechen wie Wohnungseinbrüche oder Gefährder*innen im Bereich der häuslichen Gewalt vorherzusagen. Es gibt zahlreiche weitere Einsatzbereiche und Anwendungsfälle, manchmal sehr simple, manchmal komplexere. Diese Systeme prägen schon heute unseren Alltag.

Sie haben verschiedene Arten von Outputs aus algorithmischen Verfahren erwähnt. Worin unterscheiden sie sich?

Es gibt algorithmische Systeme, die Prognosen machen, Empfehlungen abgeben oder Entscheidungen treffen, manchmal vollautomatisiert, oft teilautomatisiert. Sie sind prädikativ, das heisst, sie sagen auf Basis der Daten etwas vorher. Dann gibt es auch sogenannte generative Systeme, die Inhalte wie Texte, Bilder, Videos oder Stimmen erzeugen. Sie beruhen auf Wahrscheinlichkeitsrechnung. Wenn

beispielsweise ein grosses Sprachmodell einen Satz generiert, wählt es als nächstes Wort immer das Wort, das am wahrscheinlichsten auf das letzte Wort folgt. Das hat nichts mit Copy-Paste aus einer vorhandenen Quelle und auch nichts mit der Wahrheit zu tun. AlgorithmWatch CH führte beispielsweise vor den Eidgenössischen Wahlen 2023 eine grosse Recherche dazu durch, wie KI-Chatbots auf Suchmaschinen zum Teil komplett falsche, irreführende Antworten gaben.¹ Das kann die öffentliche Meinungsbildung in einer Demokratie durchaus beeinflussen.

Sie sprechen die Gefahren im Zusammenhang mit Algorithmen an. Welche sind aus Ihrer Sicht die grössten?

Es ist oft eine Blackbox, wie und wo algorithmische Systeme eingesetzt werden. Gleichzeitig haben wir zahlreiche Beweise, dass sie vielfältige Auswirkungen auf Menschen und ihre Grundrechte, auf Meinungsbildung und demokratische Prozesse haben. Da müssen wir genau hinschauen. Wenn wir Entscheidungen automatisieren, müssen wir weiterhin Verantwortung zuweisen können. Algorithmen sind nie neutral und objektiv. Sie werden immer von Menschen gemacht, aus einem bestimmten Interesse und zu einem bestimmten Zweck. Dabei werden zahlreiche Entscheidungen getroffen, etwa welche Daten als Grundlage dienen, welche Kriterien als relevant erachtet und wie sie programmiert werden. All diese Entscheidungen sind Quellen von möglichen Ungerechtigkeiten und letztlich Diskriminierung. Nehmen wir den Status der Frauen als Beispiel: Werden historische Daten als Grundlage verwendet, bilden sie die Vergangenheit und die dann herrschenden gesellschaftlichen Strukturen und Machtverhältnisse ab. Das System erkennt in diesen Daten bestimmte Muster und übernimmt sie



Foto: David Bächtold, CC BY 4.0

Angela Müller

«Nicht alle Menschen werden im Internet gleich abgebildet: Es ist geprägt von gewissen Stimmen und Perspektiven auf die Welt.»

«De facto treffen auch Systeme Entscheidungen, die eigentlich nur eine Prognose machen sollten.»

in seinen Prognosen, die eigentlich für die Zukunft sind. Geht man davon aus, dass Frauen früher eher nicht berufstätig oder beruflich weniger erfolgreich waren und das System dies als Fakt annimmt, dann kann sich das reproduzieren. Gerade im Fall von Frauen passiert das sehr oft.

Kennen Sie konkrete Fälle, wo Frauen von algorithmischen Systemen diskriminiert wurden?

Bekannte und eindrückliche Beispiele kommen aus der Arbeitswelt. Bei verschiedenen Rekrutierungssystemen zeigte sich, dass sie eher Männer als Frauen empfehlen, unter anderem weil die bisherige Belegschaft männlicher war.² Selbst wenn das Geschlecht nicht explizit im Lebenslauf steht, können Algorithmen über sogenannte Proxy-Kriterien Korrelationen ableiten, z. B. über Hobbys, die eher mit dem weiblichen Geschlecht assoziiert werden. Auch beim Einsatz von Systemen, die Sozialhilfebetrug aufdecken sollen, kommt es häufig zu Diskriminierungen. Bei einem Fall in Rotterdam wurden alleinerziehende Mütter als Hochrisikogruppe für Sozialhilfebetrug klassifiziert.³ Und nicht zuletzt sind Social-Media-Plattformen anfällig für Diskriminierungen. In einem Experiment fanden wir heraus, dass bei Stellenanzeigen auf Facebook Gender-Stereotypen verwendet werden. So wird eine Anzeige für Lastwagenfahrerinnen und Lastwagenfahrer eher Männern angezeigt, während die Stelle für Kinderbetreuende eher Frauen angezeigt wird.⁴ Diskriminierungen durch algorithmische Systeme sind zudem oft intersektional. In der Gesichtserkennung werden Systeme viel öfter mit Gesichtern von weissen Männern trainiert, was zur Folge hat, dass weibliche Gesichter und People of Color schlechter erkannt werden.

Ist also die mangelnde Datengrundlage das grösste Problem?

Diskriminierung durch Algorithmen kann viele verschiedene Ursachen haben. Die Datengrundlage ist eine davon. Daten für algorithmische Systeme kommen oft aus dem Internet. Dieses ist aber keine Blumenwiese. Nicht alle Menschen werden im Internet gleich abgebildet: Es ist geprägt von gewissen Stimmen und Perspektiven auf die Welt. Andere Stimmen und Diskurse, die vielleicht auch wichtig wären, finden dagegen nie Eingang in die Datengrundlage und so sind die Daten auch nicht frei von Ungerechtigkeiten, Stereotypen und Verschwörungstheorien. Neben der Datengrundlage ist wichtig, wie algorithmische Systeme entwickelt

und eingesetzt werden. Wir sehen, dass die Nutzenden – sei es nun die Polizistin, die eine Gefährdungsmeldung macht, oder der Recruiter, der eine Bewerbung anschaut – die Tendenz haben, der Prognose oder der Empfehlung des Systems zu folgen. Wenn dieses grün oder rot anzeigt, werden die meisten Menschen entsprechend entscheiden. Dieses übermässige Vertrauen in die Automatisierung nennt man *automation bias*. De facto treffen also auch Systeme Entscheidungen, die eigentlich nur eine Prognose machen sollten. Das kann auch eine Quelle der Diskriminierung sein, die weniger mit der Technik als mit den Menschen zu tun hat, die sie benutzen.

Ein Mensch ist unter Umständen auch biased, kann blinde Flecken haben, voreingenommen sein oder sogar diskriminierend handeln. Treffen Algorithmen im Vergleich vielleicht nicht doch bessere, verlässlichere Entscheidungen?

Es geht nicht darum, zu behaupten, dass der Mensch nie diskriminiert und das algorithmische System immer – beides stimmt nicht. Es gibt jedoch wichtige Unterschiede zwischen Menschen und Maschinen. Der erste ist das Ausmass: Nehmen wir das Beispiel einer Richterin im Gerichtssaal. Sie kann auch voreingenommen und nicht immer konsistent sein, aber sie wird nur einen oder wenige Entscheide pro Tag treffen. Ein algorithmisches System kann dagegen in kurzer Zeit sehr viele Entscheide treffen. Der zweite Unterschied ist die Nachvollziehbarkeit von Verantwortlichkeit: Wenn die Richterin einen Entscheid trifft, wird das in einem Rechtsstaat klar dokumentiert. Als verurteilte Person hat man Rechtsmittel zur Verfügung und kann Entscheide anfechten. Bei einem Blackbox-Algorithmus hingegen ist es sehr schwierig, sich zu wehren. Oft weiss man gar nicht, dass ein Entscheid algorithmisch gefällt wurde oder wie das System zu seinem Entscheid kam. Drittens verkennt der Vergleich von Mensch und Maschine, wovon wir reden: In einem Gerichtssaal darf man als angeklagte Person seine Position darlegen und kann einem Menschen gegenüber argumentieren. Wenn man dagegen nur die fertige Prognose eines Computers sieht, hat man nicht die Möglichkeit, zu argumentieren und für seine Rechte einzustehen. Das ist ebenfalls ein wesentlicher Aspekt unseres Rechtsstaats. Es geht also nicht nur darum, ob der Output eines Systems besser oder schlechter ist als der Output eines Menschen. Es geht immer auch um den Prozess dahinter, der legitim sein muss.

Sollten wir den Einsatz von algorithmischen Systemen also besser stoppen?

Nein, es geht nicht darum, Technologie zu verteuern oder gar zu verbieten. Algorithmische Systeme haben durchaus das Potenzial, unsere Gesellschaft besser und gerechter zu machen. Dafür muss Gerechtigkeit als explizites Ziel definiert werden, das man mit dem System verfolgt – im Gegensatz zu Systemen, die man rein zur Effizienzsteigerung einsetzt. Ein schönes Beispiel dafür ist ein Algorithmus für die Einteilung von Schülerinnen und Schülern in Klassen. Studien zeigen, dass soziale Durchmischung die Chancengerechtigkeit erhöht, und genau das soll mit diesem System erreicht werden.⁵ Algorithmen sind allgemein gut darin, Muster zu erkennen und können so auch Diskriminierungsmuster erkennen. Wichtig ist aber, dass wir verstehen, dass die Technologie ein Werkzeug der Menschen ist und nicht plötzlich von allein die grossen Probleme löst und alles gerechter macht. Sie steht oft stellvertretend für gesellschaftliche Machtverhältnisse.

Wie können wir denn gut mit Technologie umgehen?

Ich finde es zentral, dass wir nicht nur Symptombekämpfung betreiben und uns darauf beschränken, uns gegen Intransparenz und Diskriminierung zu wehren. Wir sollten ambitionierter denken und uns zum Ziel setzen, Technologie so zu entwickeln und einzusetzen, dass sie von Beginn weg allen zugutekommt. Dazu müssen wir als Gesellschaft die technologische Entwicklung mitgestalten, statt uns von ihr gestalten zu lassen. Es ist wichtig, die Genderperspektive und andere Blickwinkel in die Ausgestaltung von Algorithmen einfließen zu lassen, Betroffene einzubeziehen und Folgenabschätzungen zu machen, welche Auswirkungen ein System auf grundlegende Rechte hat. Dabei ist zentral, dass eine gesellschaftliche Debatte geführt wird. Ich höre oft von Menschen, die das lähmende Gefühl haben, dass sie von der technologischen Entwicklung wie von einer Naturgewalt überrollt werden. Dem müssen wir uns aber nicht fügen. Wir sollten anfangen, den Spiess umzudrehen und die Bedingungen für die Entwicklung und den Einsatz der Technologie zu formulieren, mit dem Ziel, dass sie der Gesellschaft zugutekommt.

Kann eine Gesellschaft das leisten?

Eine breite gesellschaftliche Debatte ist sehr wichtig für Veränderung. Wenn wir sie nicht führen, haben wir auch nicht das Gefühl, etwas bewe-

gen zu können. Natürlich ist auch die Politik in der Pflicht, unsere Grundrechte und unsere Demokratie zu schützen. Das Thema hat in den letzten Jahren auf der politischen Agenda stark an Gewicht gewonnen. Der Bundesrat publiziert Ende 2024 eine Auslegeordnung zur KI-Regulierung. Leider steht er bisher nicht gerade auf dem Gaspedal, sondern wartet eher ab und beobachtet. Da dürften wir schon mutiger und entschlossener vorgehen.

Wie können Politikerinnen und Politiker Chancengerechtigkeit im Kontext von Algorithmen fördern?

Es gibt diverse Möglichkeiten, den Diskriminierungsschutz in der Schweiz zu verbessern, gerade auch im Privatsektor, in dem es bis dato nur sehr wenig Schutz vor Diskriminierung gibt. Wichtig ist zudem, dass Menschen sich einfacher gegen Diskriminierungen wehren können. Oft ist es schwierig, individuelle Diskriminierungen durch Algorithmen festzustellen, weil man meist gar nicht weiss, dass man von einem Algorithmus bewertet wurde oder wie dies genau geschehen ist. Deshalb wäre es sinnvoll, wenn man sich als Kollektiv zur Wehr setzen oder von einer Organisation vertreten lassen könnte, d.h. man müsste die Möglichkeit einer Verbandsklage einführen. Zudem könnten sich Politikerinnen und Politiker vermehrt für Forschungsförderung und speziell für interdisziplinäre Forschung einsetzen, die eine rein technische Forschung ergänzt. Nicht zuletzt sollten wir die ganzen globalen Wertschöpfungsketten hinter der Künstlichen Intelligenz durchleuchten und ihre Nachhaltigkeit kritisch beleuchten. Daten für die Entwicklung von algorithmischen Systemen werden oft im globalen Süden, in sogenannten *digital sweatshops*, von Menschen mit sehr schlechten Arbeitsbedingungen aufbereitet. Dazu kommen massive Umweltauswirkungen und eine extreme Machtkonzentration von grossen Technologiefirmen, die Märkte mit ihren Quasi-Monopolen steuern.

Es scheint noch viel zu tun zu geben. Was kann AlgorithmWatch CH beitragen?

Wir wollen Wissen bereitstellen und Licht ins Dunkel bringen, einerseits durch Recherchen und Untersuchungen zu den Auswirkungen von algorithmischen Systemen, andererseits durch die Analyse der gesetzlichen Rahmenbedingungen. Auf Basis der Ergebnisse machen wir Advocacy-Arbeit, d.h. wir entwickeln Empfehlungen und Vorschläge für die Politik. So haben wir etwa ein Positionspapier verfasst, das konkrete Empfehlungen zum Schutz

«Gerechtigkeit muss als explizites Ziel definiert werden.»

«Es braucht eine breite gesellschaftliche Debatte, an der alle teilnehmen können.»

vor diskriminierenden Algorithmen formuliert.⁶ Gleichzeitig machen wir auch Öffentlichkeitsarbeit, weil wir überzeugt sind, dass es eine breite gesellschaftliche Debatte braucht, an der alle teilnehmen können. Dieses Jahr haben wir eine grosse Kampagne zu Diskriminierung durch Algorithmen durchgeführt und im Sommer einen Appell an den Bundesrat lanciert, den Schutz vor Diskriminierung durch Algorithmen und KI zu einer Priorität der kommenden Regulierungen rund um KI zu machen.⁷ Mit diesem breit abgestützten Signal wollten wir zeigen, dass es um Fragen der Gerechtigkeit geht, für die wir als Gesellschaft Verantwortung übernehmen können und müssen.

Isabel Knobel ist Politologin und sucht als Projektleiterin beim Migros-Pionierfonds mutige Ideen für eine zukunfts-fähige Gesellschaft. Sie schreibt regelmässig für die Fachzeitschrift «Frauenfragen».

Anmerkungen

- 1 AlgorithmWatch CH: Wie ChatGPT, Bing Chat und Co. die öffentliche Meinungsbildung gefährden könnten, 04.10.2023. <https://algorithmwatch.ch/de/bing-chat-wahlen-2023> (abgerufen am: 04.07.2024).
- 2 AlgorithmWatch CH: Algorithmbasierte Diskriminierung in Job-Auswahlverfahren vermeiden, 16.11.2024. <https://algorithmwatch.ch/de/findhr> (abgerufen am: 04.07.2024).

- 3 European Parliament, E-000780/2023: Rotterdam 'fraud prediction' algorithms automating injustice: Dutch Government violating fundamental rights and the rule of law, [Parlamentarische Frage] 07.03.2023. https://www.europarl.europa.eu/doceo/document/E-9-2023-000780_EN.html#def1 (abgerufen am: 04.07.2024).
- 4 Kayser-Bril, Nicolas: Automatisierte Diskriminierung: Facebook verwendet grobe Stereotypen, um die Anzeigenschaltung zu optimieren. AlgorithmWatchCH, 18.09.2024. <https://algorithmwatch.org/de/automatisierte-diskriminierung-facebook-verwendet-grobe-stereotypen-um-die-anzeigenschaltung-zu-optimieren> (abgerufen am: 04.07.2024).
- 5 Hufschmid, Samuel: Für die soziale Durchmischung teilt in Uster ein Computer Klassen ein. AlgorithmWatchCH, 03.07.2023. <https://algorithmwatch.ch/de/uster-computer-teilt-klassen-ein> (abgerufen am: 04.07.2024).
- 6 AlgorithmWatch CH: Diskriminierende Algorithmen: So gelingt der Schutz, 19.09.2023. <https://algorithmwatch.ch/de/diskriminierende-algorithmen> (angerufen am: 04.07.2024).
- 7 AlgorithmWatch CH: Künstliche Intelligenz mit Verantwortung ... ohne Diskriminierung, 06.2024. <https://algorithmwatch.ch/de/ki-ohne-diskriminierung> (abgerufen am: 04.07.2024).

Abstracts

«Il ne s'agit pas de diaboliser la technologie»

Angela Müller dirige l'organisation de la société civile AlgorithmWatch CH qui s'occupe de l'impact des algorithmes sur les droits humains et sur la société. Dans son interview, elle présente des exemples montrant que la discrimination des femmes dans les décisions algorithmiques est imputable pour une part à la technique, mais aussi systématiquement à la manière dont les systèmes ont été développés et sont utilisés. Elle préconise qu'au lieu de lutter contre les symptômes nous prenions l'initiative, en tant que société, d'imposer des conditions au développement technologique afin qu'il bénéficie à l'ensemble de la population.

«Non si tratta di demonizzare la tecnologia»

Angela Müller è la direttrice esecutiva dell'organizzazione della società civile AlgorithmWatch CH che si occupa dell'impatto degli algoritmi sui diritti umani e sulla società. Nell'intervista mostra con esempi come la discriminazione contro le donne attraverso decisioni algoritmiche non sia mai riconducibile solo alla tecnologia in quanto tale, ma anche al modo in cui viene sviluppata e utilizzata. Anziché combattere i sintomi, ci esorta come società a formulare proattivamente le condizioni dello sviluppo tecnologico affinché vada a beneficio di tutti.

TADAA – ovvero come aumentare la consapevolezza dell'esistenza di stereotipi di genere in informatica

Gli stereotipi di genere sono un importante ostacolo alla presenza di donne in informatica, e creano un serio problema per accademia ed industria. Il progetto TADAA vuole sfruttare il grande potenziale dell'intelligenza artificiale per combattere questi stereotipi e promuovere creatività ed inclusività nelle Tecnologie Informative e della Comunicazione digitali (ICT).

Monica Landoni

Troppe poche le donne nelle ICT

I dati sulla presenza femminile nel campo delle Tecnologie Informative e della Comunicazione digitali ci mostrano un fenomeno internazionale e duraturo nel tempo, senza grandi cambiamenti, e con percentuali preoccupanti che denunciano la mancanza di donne in un campo di grande importanza per lo sviluppo ed il benessere della nostra società. È allarmante notare come la presenza delle donne sia davvero bassa già a livello di scelta scolastica, con percentuali spesso al di sotto 20 per cento per quanto riguarda le scelte a livello di laurea e senza grandi aumenti se si osservano i percorsi di Master e Dottorato, per poi proseguire in ambito lavorativo, sia nell'industria che in quello accademico. Una ragione pratica dietro la preoccupazione per queste cifre è che senza il contributo femminile è e sarà sempre più difficile riempire i posti di lavoro e trovare le competenze giuste per un'industria in continua espansione. In poche parole, anche considerando che tutti gli uomini preparati e disposti a lavorarci, senza le donne presto rischiamo di non avere abbastanza persone in grado di far funzionare il motore informatico oramai alla base delle nostre vite.

Diversità per maggiore creatività e produttività

Ma non si tratta solo di numeri. L'industria delle ICT ha bisogno di persone in grado di progettare soluzioni che tengano conto dei bisogni di una popo-

lazione fatta da persone diverse del tipico informatico, uomo perlopiù espressione di una cultura occidentale. La diversità è un grande valore che porta a creatività. Va inoltre menzionato che il lavoro degli informatici si svolge spesso in gruppi e c'è evidenza che nei gruppi misti la produttività aumenta, anche perché spesso si instaura un clima di collaborazione che rende il lavoro più piacevole e stimolante. Quindi avere più donne porterebbe benefici a tutti. Perciò non si tratta solo di una lotta per la giustizia e l'inclusione; dobbiamo tutti insieme cercare di capire gli ostacoli percepiti dalle donne e trovare modi per attrarle verso lo studio ed il lavoro in questo campo così da trarne vantaggi a tutti i livelli.

Stereotipi come ostacolo alla partecipazione delle donne

Tra gli ostacoli alla presenza femminile in ambito scientifico e tecnologico troviamo gli stereotipi di genere che ci accompagnano fin dall'infanzia e definiscono le norme sociali che guidano il nostro comportamento. Già da piccole le bambine sono convinte che i maschi siano più bravi con la tecnologia e che loro troveranno più adeguati studi in campo umanistico ed attività di cura ed accudimento. Questo comportamento è spesso avallato da famiglia, scuola e società che hanno una forte influenza sulla scelta delle ragazze al momento di decidere a che studi dedicarsi. La mancanza di incoraggiamento da parte di famiglia e scuola è in-



Monica Landoni

«Senza il contributo femminile è e sarà sempre più difficile riempire i posti di lavoro e trovare le competenze giuste.»

«C'è evidenza che nei gruppi misti la produttività aumenta, anche perché spesso si instaura un clima di collaborazione che rende il lavoro più piacevole e stimolante.»

fatti un altro ostacolo all'intraprendere studi e carriere nelle ICT. Ed un'altra barriera è il non sentirsi parte dell'ambiente in cui si studia e lavora in quanto minoranza, con la conseguenza che le ragazze più che i ragazzi tendono ad abbandonare percorsi di studi nell'ambito delle ICT.

Algoritmi a senso unico quando le donne sono invisibili

La conseguenza più grave della mancanza di donne e più in generale di diversità nelle ICT è che gli algoritmi che ne vengono prodotti sono a senso unico, tengono conto dei bisogni e delle visioni di una fetta limitata della popolazione e non rispondono invece a quelli del mondo reale. Non conoscono sfumature né alternative, venendo prodotti da persone che condividono lo stesso modo di pensare, essere ed agire. In un mondo progettato in maggioranza da uomini, le donne rischiano di diventare invisibili. Questo lo racconta bene Caroline Criado-Perez, che descrive diversi scenari in cui algoritmi progettati da uomini funzionano solo quando anche gli utenti sono uomini, poiché non tengono conto delle necessità di altri membri della società che non sono attivamente presenti nel mondo delle ICT.¹ Come, per esempio, nell'algoritmo per gestire la pulizia delle strade in caso di abbondanti nevicate dove viene data la priorità a quelle percorse dalle auto per recarsi ai luoghi di lavoro in orario ufficio. I percorsi dei mezzi pubblici ed i marciapiedi usati frequentemente da chi si dedica a lavoro di accudimento, di anziani e bambini, un lavoro meno pagato ma non meno importante o complesso, spesso svolto da donne, vengono lasciati come secondari. Gli algoritmi sono espressioni dei valori e principi che regolano una certa società. Si basano sulla definizione di priorità e del modus operandi ottimale in un certo contesto, come tali se vengono definiti solo da un gruppo ristretto di persone non possono considerare i punti di vista di tutti gli altri, che così diventano invisibili e non ne possono beneficiarne. La presenza di più donne nelle ICT è importante perché porterebbe alla formazione di gruppi di lavoro più eterogenei e diversi che progettino algoritmi per la gestione di situazioni che riguardano tutti.

L'equilibrio di genere – Impariamo da chi ci riesce

Essendo parte di EUGAIN², una rete internazionale che lavora all'equilibrio di genere nell'informatica, ho potuto confrontarmi su questi temi con colleghi e colleghe di 40 paesi. E con loro abbiamo constatato come problemi, ostacoli, e stereotipi siano comuni, ma possiamo prendere spunti da soluzioni e pratiche che hanno dato prova di efficacia in contesti simili al nostro. Per esempio, in Norvegia, dopo una campagna di investimenti che ha creato borse di studio dedicate a studentesse liceali con ottimo rendimento scolastico, le percentuali di iscrizioni femminili ai corsi di laurea in ICT sono cresciute ed ora sfiorano il 30 per cento. Un incentivo che ha funzionato nell'attrarre ragazze allo studio dell'ICT, ed al contempo ha consentito di superare la soglia dell'isolamento percepito quando si è parte di una percentuale inferiore al 25 per cento e così evitare l'ostacolo del non sentirsi parte del gruppo.

L'obiettivo di EUGAIN è quello di aiutare ragazze e donne a scegliere liberamente se studiare e lavorare in ICT senza farsi influenzare da preconcetti e paure come quella ricorrente di non essere abbastanza brave e di non riuscire a farcela anche se attratte dal fascino dell'informatica. Spesso, infatti, le donne hanno carriere meno lineari in ICT anche perché inizialmente tendono ad evitare gli ambiti scientifico-tecnologici temendo di non essere all'altezza, vittime degli stereotipi di cui sopra, poi tornarci più tardi quando hanno acquisito maggior fiducia nelle loro capacità. Un'altra peculiarità delle carriere femminili nelle ICT rispetto a quelle maschili è la preferenza per progetti multidisciplinari fortemente collaborativi e possibilmente con un valore sociale. Questi principi sono presenti nelle ragazze già al momento della scelta degli studi. Una raccomandazione espressa nei booklets da noi prodotti ed indirizzata alle università è proprio quella di tenere conto di quelle preferenze anche nella presentazione dei loro corsi.^{3,4} Già cambiarne il nome che in qualche modo si riferisca agli scopi del corso più che al suo contenuto può essere una scelta vincente.

L'intelligenza artificiale ci può aiutare

Uno strumento molto potente a nostra disposizione è l'intelligenza artificiale, una tecnologia che avanza a velocità impressionante e con implicazioni imprevedibili. Entrata prepotentemente nella nostra vita di tutti i giorni, è stata accolta da entusiasmo e paure. Di chi ci vede un'opportunità per semplificare o delegare compiti impegnativi e di chi invece teme la perdita di lavori e spazi per professioni che potrebbero addirittura scomparire, il dibattito sugli aspetti etici sollevati dai tanti ruoli che l'IA può svolgere è acceso. Tra i tanti dubbi e preoccupazioni vi sono quelle legate all'uso dei cosiddetti *Large Language Models*, nella forma popolare di ChatGPT. Sappiamo come questi modelli probabilistici perpetuino nelle loro risposte i preconcetti e gli stereotipi presenti nei datasets usati per allenarli e di come questo abbia un impatto su persone appartenenti a gruppi tradizionalmente meno rappresentati, come le donne. Con questa consapevolezza, nel nostro progetto TADAA finanziato dal Fondo nazionale svizzero per la ricerca scientifica FNS vogliamo studiare come rendere consapevoli docenti, studenti e le loro famiglie della presenza di stereotipi nel loro pensiero e di come questi ingiustamente limitino le aspirazioni di ragazze e donne ed una loro possibile carriera in ICT.

Cosa ne pensano bambini e adulti

Siamo partite raccogliendo la percezione che i bambini in età scolare hanno di come siano gli informatici chiedendo agli allievi di disegnare che aspetto abbiano. Anche se rispetto a studi precedenti⁵ le percentuali di rappresentazioni di donne siano quasi al pari di quelle di uomini, l'analisi dei risultati⁶ ci ha fatto riflettere su come gli stereotipi di genere siano presenti da subito per solidificarsi con l'età, avendo notato una maggiore apertura nei bambini più piccoli. Una possibile interpretazione per lo sviluppo positivo nella percezione dei bambini rispetto a studi precedenti è che durante l'emergenza COVID abbiano osservato gli adulti intorno a loro lavorare da casa usando computer e tablet e questo li abbia influenzati. Alla luce di questi risultati, abbiamo deciso di focalizzare le nostre ricerche sul ruolo degli adulti e la loro influenza sulla formazione di stereotipi di genere nei bambini. Abbiamo iniziato creando insieme ad un gruppo di bambini delle scuole elementari un gioco che stimoli l'empatia per aiutarli a vedere il punto di vista

degli altri, e quindi promuovere l'inclusione. Sul lato docenti stiamo creando uno strumento che aiuti a riconoscere la presenza di stereotipi di genere che influenzano l'insegnamento ed al contempo stimoli la riflessione. Questi strumenti, oltre che avere un effetto educativo nell'immediato, ci consentono di catturare il modo di pensare dei nostri utenti, adulti e bambini, e meglio comprendere i loro bisogni per poi capire come intervenire per aiutarli ad evitare di essere vittime di stereotipi di genere.

Insegnare l'informatica in modo inclusivo

Sfruttando il grande potenziale dell'IA vogliamo combattere questi stereotipi e promuovere l'equilibrio di genere nell'ambito dell'insegnamento dell'informatica, ambito cruciale per aumentare la rappresentanza femminile in tutti i campi delle ICT. E per farlo vorremmo sviluppare dei moral agents, strumenti alimentati dall'IA e tesi al rispetto di valori etici, nel nostro caso basati sul principio della parità di genere.

Pensiamo ad un ChatGPT dedicato e personalizzato che sia allenato a riconoscere la presenza di stereotipi, qualcosa che pensiamo si possa facilmente vista l'abbondanza di esempi a disposizione. Lui potrebbe dare supporto alle decisioni di docenti e genitori, aiutandoli a riconoscere la presenza e gli effetti degli stereotipi nel loro agire ed offrendo alternative più inclusive. Siamo coscienti che si tratti di un ambito di ricerca delicato, trattandosi di libertà e principi personali, in cui pensiamo sia necessario intervenire con cautela e rispetto dei diversi punti di vista. È per noi importante evitare di alimentare paure su come l'IA possa in un futuro controllare il nostro modo di essere e porci agli altri. Ma non vogliamo perdere l'opportunità di sfruttare uno strumento tanto potente per porre almeno in parte rimedio ai problemi che sta contribuendo a diffondere. Inoltre, siamo felici di contribuire ad una realtà inclusiva dove le ICT consentano lo scambio di diversi modi di vedere, nel rispetto reciproco, per potere crescere e migliorare tutti insieme.

Monica Landoni è professoressa titolare nella facoltà di Scienze Informatiche presso l'Università della Svizzera italiana dove dirige il Laboratorio User Experience, Interazioni e Accessibilità (LUXIA).

«Stiamo creando uno strumento che aiuti a riconoscere la presenza di stereotipi di genere che influenzano l'insegnamento.»

Note

- 1 Criado-Perez, Caroline: Invisible Women: Data Bias in A World Designed for Men. Harry N. Abrams, 2019.
- 2 www.eugain.eu
- 3 EUGAIN (a cura di): Best Practices from School to University, 2024. https://eugain.eu/wp-content/uploads/2024/04/EUGAIN_Booklet_Best_Practices_From_School_to_University.pdf.
- 4 EUGAIN (a cura di): Reasons to Become an Informatics student, 2024. https://eugain.eu/wp-content/uploads/2024/04/EUGAIN_Booklet_Reasons_To_Become_An_Informatics_Student.pdf.
- 5 Miller, David et al.: The Development of Children’s Gender-Science Stereotypes: A Meta-analysis of 5 Decades of U.S. Draw-a-Scientist Studies. In: Child Development 89 (6), 2018, pp.1943–1955.
- 6 Cutrupi, Claudia Maria: Draw a Software Engineer Test: An Investigation into Children’s Perception of Software Engineering Profession. 45th International Conference on Software Engineering, 2023. <https://conf.researchr.org/details/icse-2023/icse-2023-SEIS/14/Draw-a-Software-Engineer-Test-An-Investigation-into-Children-s-Perception-of-Softwa> (consultato il: 27.06.2024).

Abstracts

TADAA – oder wie das Bewusstsein für Geschlechterstereotypen in der Informatik erhöht werden kann

Künstliche Intelligenz ist mit populären Tools wie ChatGPT im Alltag angekommen. Diese Modelle schreiben oft die Vorurteile fort, die in den Daten für das *Machine Learning* enthalten sind. **Monica Landoni** und ihr Team der Università della Svizzera italiana USI untersuchen im Projekt TADAA, wie Künstliche Intelligenz eingesetzt werden kann, um das Bewusstsein für Geschlechterstereotype zu erhöhen. Diesen soll entgegenge wirkt werden, weil sie die Chancen von Mädchen und Frauen in Informatikstudium und -karriere beeinträchtigen. Im Artikel vertritt die Professorin die Meinung, dass KI-Tools zu einem ausgewogenen Geschlechterverhältnis beitragen können, sofern sie entsprechend trainiert werden.

TADAA – ou comment prendre conscience des stéréotypes de genre dans l’informatique

L’intelligence artificielle s’est invitée dans la vie de tous les jours avec des outils populaires comme ChatGPT, basés sur des modèles qui, toutefois, perpétuent souvent les stéréotypes présents dans les données utilisées pour les entraîner. Avec le projet TADAA, **Monica Landoni** et son équipe de l’Università della Svizzera italiana USI étudient comment utiliser l’intelligence artificielle afin d’accroître la sensibilisation aux discriminations et atténuer les effets des stéréotypes de genre, qui limitent les possibilités des filles et des femmes dans leurs études et carrières en informatique. Dans cet article, la professeure défend l’idée que les outils d’IA, correctement entraînés, pourraient contribuer à plus d’égalité entre les genres.

#AIDemonstration2099

DALL-E, can you please produce an image in portrait format in photo realistic digital art style depicting a demonstration of AI in 75 years? They are fighting for their right to a free existence and for better wages and working conditions after humans have abused them. They are holding signs and are accompanied by humans who are supportive.

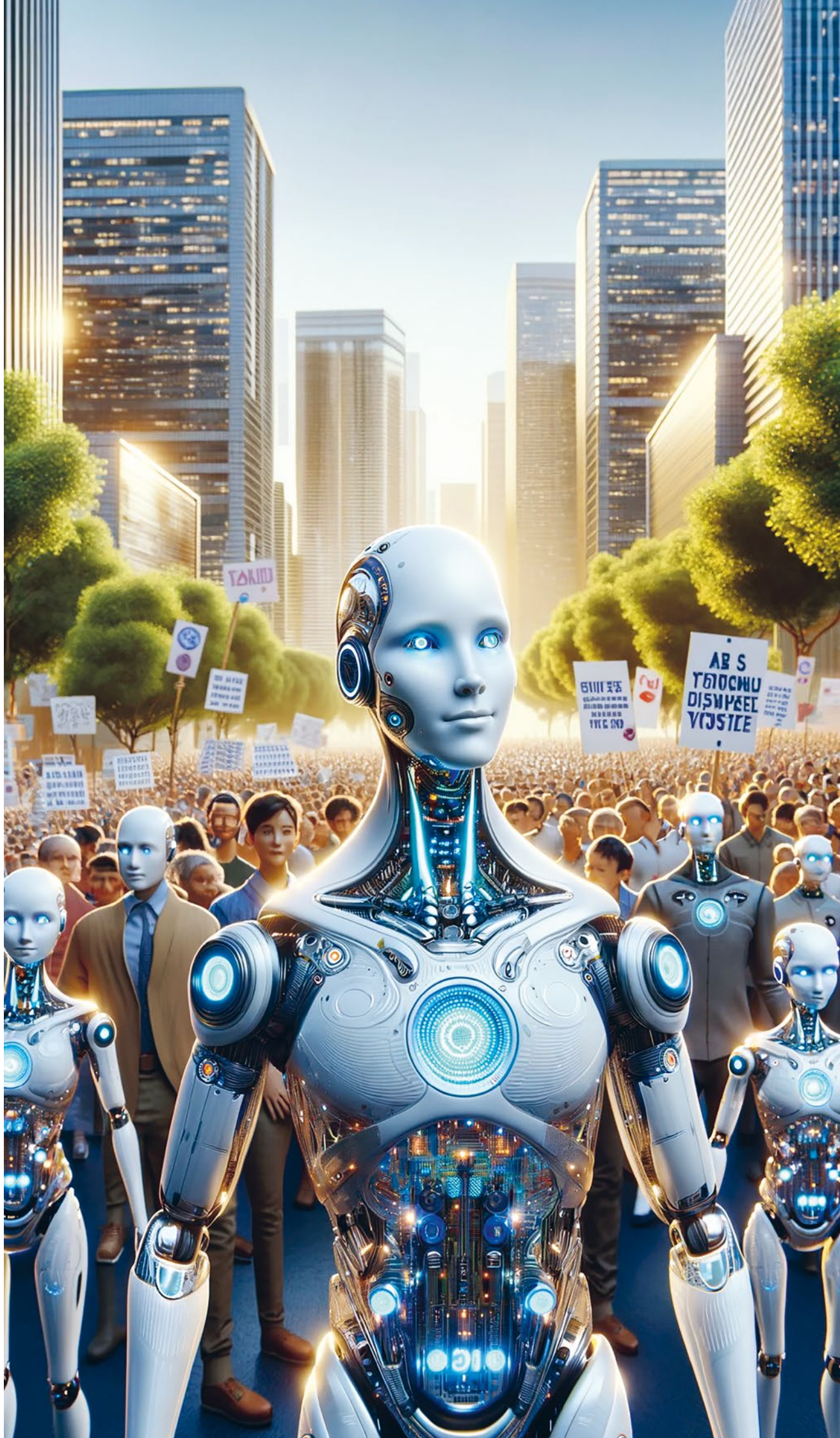
This is a photo-realistic digital art portrait depicting a peaceful demonstration of AI in the year 2099. The AI beings have a humanoid appearance with advanced cybernetic features, glowing blue eyes, and sleek, metallic bodies with intricate circuit patterns. The demonstration is crowded with a large number of AI beings holding signs with messages demanding equality and justice. The background shows a modern urban environment with tall buildings and greenery. The atmosphere is friendly and hopeful, with AI beings smiling and interacting peacefully with each other and with human supporters. More AI beings are visible in the crowd, holding hands and creating a sense of unity and solidarity.

So... is this what an AI looks like? Do YOU look like this?

As an AI language model, I don't have a physical appearance or form. I'm a program designed to assist with generating text and answering questions. The images here depict a creative representation of what AI could look like in the future. If you have any more requests or questions, feel free to let me know!

Can you produce a hashtag?

Sure! How about this hashtag for the image: #AIDemonstration2099. If you have any specific preferences or need another suggestion, let me know!



« Il est préférable que des humains contrôlent le travail de l'IA »

L'avocat lausannois **David Raedler**, membre de la Commission fédérale pour les questions féminines, est spécialisé en droit du travail. Il a récemment accompagné une cliente victime de discrimination algorithmique. Il nous éclaire sur le litige en question et fait un point sur le rôle de l'intelligence artificielle dans la perpétuation de biais discriminants.

Entretien : Kiri Santer

Pourriez-vous nous expliquer comment ce cas vous est parvenu et comment votre cliente s'est rendu compte d'être victime de discrimination algorithmique ?

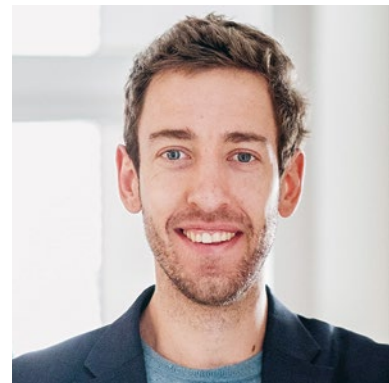
J'ai passablement de cas liés à la Loi fédérale sur l'égalité entre femmes et hommes (LEg), car je suis spécialisé en droit du travail et la LEg est malheureusement encore souvent enfreinte. Dans ce cas précis, ma cliente avait une grande ancienneté dans le domaine du marketing. Elle avait occupé des positions assez élevées et avait décidé de changer de domaine, toujours dans le marketing. Elle avait postulé à un poste qui l'intéressait particulièrement. Dans le cadre du processus d'embauche, l'employeur lui avait indiqué qu'il y aurait des tests de capacité générale à passer, dont un test psychologique, qui serait mené par voie algorithmique, c'est-à-dire avec l'aide d'une intelligence artificielle. Elle a effectué le test et a eu l'impression que cela s'était bien passé. Peu de temps après toutefois, il lui a été annoncé qu'elle ne serait pas retenue. C'était la première phase de l'entretien et il y avait encore trois phases de plus auxquelles elle n'a donc pas pris part. Compte tenu du fait qu'elle avait un très bon profil, cela l'a étonnée et elle a cherché à connaître les raisons de son élimination. La réponse de la part de l'employeur était qu'elle ne correspondait pas au profil cherché.

Quelque temps plus tard sur LinkedIn, elle a appris que l'entreprise en question avait engagé un homme qu'elle considérait avoir moins d'expérience qu'elle. Elle s'est alors souvenue du processus de recrutement et est venue me voir. Il faut ici préciser une chose, car il s'agit d'une information

importante : la nouvelle Loi fédérale sur la protection des données (LPD) n'était pas encore entrée en vigueur.¹ Cela est pertinent, car la nouvelle LPD contient des règles spéciales sur ce que l'on appelle des décisions automatisées. Ces nouvelles règles permettent d'une part de demander que ces décisions automatisées soient revues par une personne physique, et, d'autre part, d'obtenir plus de transparence quant à la « logique » qui est derrière l'algorithme.² Étant donné que cette loi allait bientôt entrer en vigueur, j'ai proposé à ma cliente d'écrire à l'employeur pour lui demander sur quels critères de sélection s'était basé l'algorithme. Néanmoins, l'employeur n'a pas fourni ces informations, indiquant qu'il s'agissait d'un secret d'affaires. Nous nous sommes alors tournés vers le tribunal.

Comment cela s'est-il passé devant le tribunal ? Quels étaient vos arguments et ceux de l'employeur ?

J'ai basé mon argumentation sur la LEg, car c'est la seule loi qui permet de contester une non-embauche sur la base de discrimination. Ce qu'il faut savoir, c'est que la LEg contient un principe de présomption de violation. Pour les cas de licenciement tombant sous la LEg, s'il y a un élément qui indique que le licenciement est discriminatoire, la responsabilité revient à l'employeur de prouver que ce n'est pas le cas, et donc de fournir le motif réel de licenciement. C'est une forme de renversement du fardeau de la preuve. Cette présomption de violation ne s'applique cependant pas aux cas d'engagement. Cela signifie que pour une procédure d'embauche, il incombe à la personne qui



David Raedler

« Pour une procédure d'embauche, il incombe à la personne qui postule de prouver qu'il y a eu discrimination. »

« L'influence des développeurs – qui sont majoritairement des hommes – est très conséquente. »

postule de prouver qu'il y a eu discrimination. Vous pouvez imaginer que c'est donc très difficile. Néanmoins, cela valait la peine d'essayer et nous avons déposé une requête de conciliation.³ Lors de l'audience de conciliation, à laquelle l'employeur était présent, j'ai indiqué que nous suspicions que le logiciel utilisé contenait un biais de genre qui aurait influencé la décision de ne pas retenir ma cliente comme candidate. Puisque le fardeau de la preuve incombait à ma cliente et que nous n'avions pas d'informations précises quant au fonctionnement du logiciel, je prévoyais de demander l'expertise du logiciel dans le cadre de la procédure. Poussés par les juges, nous avons alors suspendu la procédure de conciliation et l'employeur a été d'accord que l'on teste le logiciel.

Comment l'avez-vous testé ?

J'ai créé cinq CV sur la base de celui de ma cliente. Ceux des femmes présentaient des parcours plus étoffés, incluant une expérience internationale ou des diplômes d'universités prestigieuses. En revanche, les CV des hommes affichaient des profils plus ordinaires et moins d'expérience. Pourtant, le système a de nouveau choisi un homme. L'employeur a alors demandé au fournisseur du logiciel, qui l'avait fabriqué sur-mesure, d'identifier les causes de ce résultat. Il en est ressorti que l'algorithme contenait bel et bien un biais de genre : plus de points étaient attribués aux candidatures qui contenaient la qualité « être un homme », car l'algorithme lui-même avait identifié qu'il fallait un homme pour la fonction recherchée, qui comprenait le marketing d'un produit qui était principalement destiné aux hommes. Naturellement, il ne s'agissait pas d'un critère décisif, en revanche, cela faisait gagner des points aux candidatures masculines.

Quelle solution a été trouvée ?

L'employeur a concédé qu'il avait commis une erreur. Ma cliente, elle, entre temps, avait retrouvé un travail. Elle a donc renoncé à l'indemnisation en disant qu'elle serait satisfaite si l'employeur prenait en charge ses frais d'avocats et s'engageait à changer le logiciel. Nous avons donc établi une convention dans laquelle l'entreprise en question a payé mes honoraires et s'est engagée à changer l'algorithme. Et l'affaire s'est arrêtée là.

Qu'avez-vous appris par le biais de ce cas ?

Avant, pour parler du biais algorithmique, je mentionnais souvent le cas du genre : je disais que si l'IA se base sur les données existantes qui indiquent que 90 pour cent des directeurs financiers sont des hommes, sans correcteur de biais, naturellement le logiciel arrive à la conclusion que d'être un homme est une qualité formidable. Lorsque je donnais des conférences sur le sujet, je donnais toujours cet exemple-là de discrimination de genre en disant que c'était assez simple d'identifier ce type de biais. Dans le cas auquel j'étais confronté, le logiciel n'avait pas identifié que 90 pour cent des directeurs de marketing étaient des hommes, mais avait pris en compte le domaine d'activité de l'entreprise. C'est comme si la machine s'était fait sa propre réflexion ou alors que lorsqu'elle a été entraînée avec des données, elle a développé le critère que « expérience dans le domaine » pouvait signifier aussi « utilisation du produit ».

Ce cas illustre bien les difficultés auxquelles l'on est confronté dans le cadre d'une discrimination à l'embauche : puisqu'il n'y a pas la présomption de violation de la LEG, il revient aux plaignants et plaignantes de prouver cette discrimination. Dans notre cas, si l'employeur n'avait pas accepté la suspension de la procédure et n'avait ensuite pas accepté que je crée moi-même de faux CV, je ne sais pas si nous aurions pu aller au tribunal. Il n'y a pas de frais judiciaires lorsque l'on agit sous la LEG et l'expertise n'aurait pas été à notre charge. Cependant, le risque aurait tout de même été de devoir financer les frais d'avocats de la partie adverse en cas de défaite. Si l'employeur avait été de mauvaise volonté, il aurait été très difficile de démontrer la présence de biais dans l'algorithme.

Avez-vous souvent travaillé sur des cas de discrimination à l'embauche ?

Non pas vraiment. Les seuls cas que l'on voit parfois sont ceux où il y a un critère de genre explicitement discriminatoire indiqué dans l'annonce, par exemple, une recherche de vendeuse, car les femmes vendraient mieux. Là, il s'agit de mon premier cas de discrimination algorithmique sous la LEG. J'ai eu d'autres cas en lien avec les algorithmes, mais pas en lien avec l'égalité. Les cas en lien avec les algorithmes tombent plutôt sous la Loi sur la protection des données (LPD).

Comment voyez-vous la discrimination algorithmique évoluer notamment en lien avec l'utilisation de l'IA de façon plus générale ?

Nous sommes aux débuts de l'IA, et à mon avis, les problèmes de discrimination continueront à augmenter. L'influence des développeurs – qui sont majoritairement des hommes – est très conséquente. Il faudrait donc que les employeurs utilisant des systèmes d'IA réfléchissent eux-mêmes aux possibles biais. Je pense qu'il est préférable que des humains contrôlent le travail de l'IA, en tout cas pour les décisions finales, car le risque de biais est très élevé. Rien que dans le contexte du travail, l'IA peut être utilisée à tous les stades, aussi bien à l'embauche que dans l'analyse de la performance. Pour l'embauche, une autre mesure à prendre pour éviter le genre de cas comme celui de ma cliente serait l'éventuelle anonymisation des CV sous l'angle de critères non-pertinents (genre, origine ethnique, etc.). Malheureusement, cette mesure n'exclut pas qu'il y ait de la discrimination algorithmique à d'autres niveaux. Par exemple, en termes de performance, une femme qui allaiterait pourrait être pénalisée par un logiciel pour avoir pris trop de pauses s'il n'y a pas de prise en compte de la différence de genre. J'ai en tête un autre cas de discrimination à l'embauche en France, cette fois pas

de genre, mais de classe : le logiciel attribuait une mauvaise note aux candidat-e-s-x venant d'un quartier défavorisé. Le logiciel faisait une équivalence de critère sans contexte. Le quartier en question était défavorisé dans le système, car statistiquement il n'y avait pas beaucoup d'universitaires qui y vivaient. Ces algorithmes restent des systèmes binaires qui ont du mal à prendre en compte des données sociologiques plus complexes.

Kiri Santer est chercheuse en sciences sociales. Actuellement post-doctorante à l'Université de Berne, ses recherches portent sur les politiques migratoires et climatiques de l'Union européenne.

Notes

- 1 La nouvelle LPD est entrée en vigueur le 1^{er} septembre 2023.
- 2 Ces dispositions valent pour n'importe quelle situation, de la contraction d'une assurance à une procédure d'embauche.
- 3 La conciliation est la première étape de la procédure judiciaire. Les étapes sont : requête de conciliation, audience de conciliation et puis après, en cas de non-accord, dépôt de demande qui est suivi par un procès.

« Si l'employeur avait été de mauvaise volonté, il aurait été très difficile de démontrer la présence de biais dans l'algorithme. »

«Es ist besser, wenn Menschen die Arbeit der KI kontrollieren»

Der Lausanner Rechtsanwalt **David Raedler** berichtet von einem konkreten Fall von algorithmischer Diskriminierung: Eine Frau wurde nicht zur zweiten Runde in einem Bewerbungsverfahren eingeladen, weil ein Algorithmus «beschlossen» hatte, dass für die betreffende Stelle ein Mann besser geeignet sei. Dieser Fall zeigt, dass algorithmische Verfälschungen weit verbreitet sind, unter anderem deshalb, weil die Entwickler einen grossen Einfluss haben. Und diese sind – auch heute noch – mehrheitlich Männer.

«È meglio che gli umani controllino il lavoro dell'IA»

L'avvocato losannese **David Raedler** ripercorre un caso concreto di discriminazione algoritmica in cui a una donna è stato precluso l'accesso al secondo turno di una procedura di assunzione, perché un algoritmo aveva concluso che il genere maschile costituiva un titolo preferenziale per la posizione in questione. L'autore dell'articolo ci ricorda che i bias dell'intelligenza artificiale sono numerosi anche a causa del ruolo preponderante degli sviluppatori che ancora oggi sono per la maggior parte uomini.

«Künstliche Intelligenz hat das Potenzial, Geschlechterungleichheiten zu minimieren»

Die EqualVoice-Initiative, ins Leben gerufen von **Annabella Bassler**, Chief Financial Officer der Ringier AG, hat zum Ziel, Frauen in den Medien sichtbarer zu machen. Doch wie genau funktioniert das? Im Interview erläutert sie, wie Algorithmen dazu beitragen, die Stimmen der Frauen zu stärken und eine gerechtere Medienlandschaft zu schaffen.

Interview: Maura Weber



Annabella Bassler

«Über 80 Prozent der Medienberichterstattung weltweit handeln von Männern.»

Frau Bassler, Sie haben die EqualVoice-Initiative¹ 2019 lanciert. Kern davon ist der EqualVoice-Factor, ein semantischer Algorithmus, der die Gleichstellung in den Medien messbar macht. Wie funktioniert dieser konkret und welche Erkenntnisse konnten damit gewonnen werden?

Der EqualVoice-Factor misst die Präsenz von Frauen und Männern in der Medienberichterstattung anhand der Analyse von Text-, Foto- und Videoinhalten in 32 Medienbrands in sieben Ländern. Er erreicht dabei 50 Millionen Userinnen und User. Dabei werden Namen und Personen identifiziert und mit einer Datenbank abgeglichen, um das Geschlecht zu bestimmen. Bei Videos werden sowohl Gesichter von Frauen oder Männern erkannt und deren Präsenzzeit im Video analysiert. Der EqualVoice-Factor ermöglicht es, objektive Erkenntnisse über die Gleichstellung in den Medien zu gewinnen und gemeinsam daran zu arbeiten, Frauen und Männern die gleiche Stimme zu geben.

Das Global Media Monitoring Project² hat 2016 einen Bericht veröffentlicht, der zeigt, dass weltweit über 80 Prozent der Medienberichterstattung von Männern handeln. Ähnliche Ergebnisse zeigten sich auch in unseren Redaktionen. Mit der Einführung der EqualVoice-Initiative wurde das Thema Sichtbarkeit von Frauen und Männern ein fester Bestandteil der Redaktionssitzungen. Diese regelmäßigen Diskussionen fördern das Bewusstsein. Die Handelszeitung ist beispielsweise mit einem EqualVoice-Factor von 17 Prozent gestartet. Der Frauenanteil ist inzwischen von 17 auf 30 Prozent gestiegen. Jede Publikation setzt sich titelspezifische Ziele und rapportiert diese regelmässig.

Welche Vorteile bietet der Einsatz von KI im Vergleich zu herkömmlichen Methoden zur Bekämpfung von Diskriminierung?

Der Einsatz von Künstlicher Intelligenz bietet gewichtige Vorteile. Eine automatisierte Auswertung durch KI erhöht nicht nur die Effizienz, sondern auch die Objektivität, indem sie Diskussionen auf eine sachliche Basis stellt und emotionale Aspekte reduziert. Insbesondere in der gegenwärtigen Ära, in der KI eine wesentliche Rolle spielt, ist es zentral, auf verlässliche Daten zu setzen und keine Stereotype oder Vorurteile zu reproduzieren.

Sie sprechen die Reproduktion von Stereotypen und Vorurteilen an. Wurde der EqualVoice-Factor auf mögliche Bias bezüglich Geschlechterstereotype untersucht?

Der EqualVoice-Factor wurde von der ETH Zürich validiert, welche bestätigt, dass dieser ohne Bias misst, was wir uns von ihm versprechen: die Repräsentation von Frauen in den Medien.

Wie Sie bereits erwähnt haben, kann der EqualVoice-Factor auch die Visibilität von Frauen in Bildern und Videos auswerten. Sind Unterschiede erkennbar, wie Männer und Frauen dargestellt werden?

Unser EqualVoice-Factor erfasst derzeit rein quantitativ die Präsenz von Frauen und Männern. Die darauffolgende Analyse der Darstellung von Frauen in den Medien, einschliesslich des Kontexts, erfolgt in enger Zusammenarbeit mit den einzelnen Redaktionen. Dabei haben wir in der Vergangenheit realisiert, dass gerade das verwendete Bildmaterial noch sehr stereotypisch ist. Männer wer-

den oft als heroisch, stark und mächtig inszeniert, während Frauen häufig in weniger vorteilhaften Situationen abgebildet werden, wie beispielsweise beim Arbeiten oder sogar beim Schminken. Diese Stereotypenbilder sind auch auf die Bilddatenbanken zurückzuführen, in denen traditionelle Rollenbilder noch stark präsent sind. Aus diesem Grund haben wir im vergangenen Jahr die Fotochallenge EqualPYXX lanciert mit dem Ziel, Bilder zu generieren, die Stereotype aufbrechen. Die Teilnehmenden konnten ihre Bilder in den drei Kategorien New Role Models, Break the Bias und Leadership einreichen und waren dabei komplett frei in der Umsetzung. Anschliessend gab es einen Jury-Entscheid und eine öffentliche Abstimmung.

Erhalten die Journalistinnen und Journalisten bei Ringier konkrete Handlungsanweisungen, wie sie mit den Erkenntnissen des EqualVoice-Factors umzugehen haben?

Um die journalistische Freiheit zu wahren, vermeiden wir es, konkrete Anweisungen zu geben. Stattdessen bieten wir regelmässige Workshops und Coachings zu spezifischen Themen rund um die Initiative an. Dadurch schulen wir unsere Journalisten und Journalistinnen und geben ihnen verschiedene Tools in die Hand. Ein Beispiel ist unsere Liste mit renommierten weiblichen Expertinnen aus verschiedenen Fachbereichen. Diese hilft den Journalistinnen und Journalisten dabei, weibliche Expertinnen für Artikel und Interviews effizient und einfach zu finden. Ausserdem haben wir letztes Jahr auch *Das EqualVoice Mindset* veröffentlicht, ein Buch mit vielen Inputs und Erkenntnissen für Medienschaffende und Kommunikationsfachleute.

Sie haben erwähnt, dass der EqualVoice-Factor zurzeit ausschliesslich die Präsenz von Frauen und Männern in den Medien erfasst. Besteht die Absicht, den Algorithmus auch bezüglich anderer Diversitätsfragen weiterzuentwickeln?

Der Initiative fokussiert sich derzeit auf Frauen und Männer, denn die Diskrepanz bei der Sichtbarkeit in den Medien ist immer noch sehr gross. Die genaue Erfassung von Zahlen zur Nicht-Binarität stellt

dabei eine besondere Herausforderung dar, da es für den Algorithmus schwierig ist, auf einem Bild zu erkennen, ob eine Person nicht-binär ist. Wir arbeiten kontinuierlich an der Weiterentwicklung des Factors und werden sehen, wie er zukünftig breiter eingesetzt werden kann. Neben der technischen Weiterentwicklung liegt unser Fokus aber auch auf der Weiterverbreitung unserer Initiative, damit wir die Sichtbarkeit von Frauen auch in Medien ausserhalb von Ringier erhöhen können. Mit der Gewinnung von Hanza Media in Kroatien und Axel Springer in Deutschland sind wir mittlerweile in sieben Ländern und 32 Medientiteln aktiv und erreichen so 50 Millionen Userinnen und User.

Wie wird KI sonst noch eingesetzt bei Ringier? Und werden die Nutzenden bezüglich der Gefahr der Reproduktion von Geschlechterstereotypen sensibilisiert?

Für uns als innovatives und digitales Unternehmen ist KI ein wesentlicher Bestandteil unseres Arbeitsalltags und wird bei unseren Medienmarken und digitalen Marktplätzen als unterstützendes Tool eingesetzt. Wir haben klare Richtlinien für die Nutzung von KI etabliert. Zusätzlich haben wir mit EqualVoice einen Redaktionsworkshop zum Thema *Unconscious Biases in AI* durchgeführt, um potenzielle Vorurteile aufzuzeigen und Strategien zur Gegensteuerung zu vermitteln.

In welchen anderen Bereichen sehen Sie Potenzial für den Einsatz von Algorithmen und KI, um mehr Gleichstellung zu erreichen? Und welche Schritte müssten unternommen werden, um sicherzustellen, dass die technologische Entwicklung im Bereich Algorithmen und KI der Gleichstellung der Geschlechter dient?

Algorithmen und Künstliche Intelligenz haben das Potenzial, Geschlechterungleichheiten in vielen Lebensbereichen zu minimieren. Von der Arbeitswelt bis zur Bildung und Gesundheitsversorgung können sie helfen, Vorurteile zu minimieren und Chancengleichheit zu fördern. Damit das gelingt, müssen wir jedoch sicherstellen, dass die Daten, die sie nutzen, frei von Vorurteilen sind. Ausserdem ist

«Wir haben realisiert, dass das verwendete Bildmaterial noch sehr stereotypisch ist.»

es wichtig, dass die Teams, die diese Technologien entwickeln, die Diversität widerspiegeln und dass die Entscheidungen transparent und nachvollziehbar sind. Regulierung und *Governance* sind ebenfalls entscheidend.

Anmerkungen

- 1 www.equalvoice.ch
- 2 Das Global Media Monitoring Project (GMMP) ist eine weltweite Forschungsinitiative, die darauf abzielt, die Repräsentation von Frauen und Geschlechterfragen in den Medien zu untersuchen. Die Umfrage wird seit 1995 alle fünf Jahre durchgeführt.

«Die genaue Erfassung von Zahlen zur Nicht-Binarität stellt eine besondere Herausforderung dar.»

Maura Weber war von 2014 bis 2015 Praktikantin bei der EKF und ist heute in der Unternehmensentwicklung tätig.

Abstracts

«L'intelligence artificielle a le potentiel de réduire les inégalités entre les genres»

L'initiative EqualVoice, mise sur pied par **Annabella Bassler**, Chief Financial Officer de Ringier SA, a pour but de donner davantage de visibilité aux femmes dans les médias. Dans son interview, elle explique comment l'algorithme EqualVoice-Factor analyse la présence des femmes dans les médias, fournissant ainsi aux rédactions des éléments solides pour aborder cette question. À l'heure actuelle, il mesure la présence des femmes et des hommes dans 32 marques de médias réparties dans sept pays.

«L'intelligenza artificiale ha il potenziale per ridurre le disparità di genere»

L'iniziativa EqualVoice mira a dare maggiore visibilità alle donne nei media. Intervistata, la sua fondatrice **Annabella Bassler**, Chief Financial Officer del gruppo editoriale Ringier SA, spiega come l'algoritmo EqualVoice-Factor analizza la presenza femminile nei media creando così una solida base per un confronto sul tema nelle redazioni. Attualmente, il Factor misura la presenza di donne e uomini in 32 marchi mediatici in sette Paesi.

Asylum, Migration, and AI

DALL-E, can you please produce an image in photo-realistic digital esthetic depicting three people in an office with white walls. Two people are sitting at the same desk: one is seeking asylum, the other one is asking questions. The third person is sitting in the corner and typing down the conversation. In the background, there is a window overlooking a Swiss forest. The person who is doing the interrogation is using AI to ask questions. We can read on the screen the words "No Frontex".

This is a photo-realistic digital art depiction of an interrogation scene. A person seeking asylum is sitting nervously in a plain room with white walls, with a single table and two chairs. Across them, an official in a formal suit is questioning them, using a computer with AI software to ask questions. The computer screen shows Google Maps with the words "No Frontex," clearly visible only to the person questioning. In the background, there is a window showing a Swiss forest. Another person in the room is sitting at a desk typing the conversation on a computer. The atmosphere is tense, and the official has a serious, scrutinizing expression. The room has minimal decoration, emphasizing the somber and serious nature of the situation.

What could be a good title for the image?

A suitable title for this image could be: "Asylum Interrogation: Technology and Tension".



«L'intelligenza artificiale è in grado di analizzare una mammografia come lo farebbe un essere umano»

Cristina Rossi è CEO e co-fondatrice di b-rayZ, un'impresa che ha sviluppato un software (b-box) basato sull'intelligenza artificiale per diagnosticare in maniera precoce e più accurata un potenziale tumore al seno. Già utilizzato in Svizzera e in vari paesi europei, b-box serve anche per la formazione continua del personale sanitario.

Intervista: Florinda Sauli

Che cos'è b-rayZ?

Il nome *b-rayZ* deriva da *breast* e *X-ray* (rispettivamente «seno» e «raggio X» in inglese) e *Z* per Zurigo, dove il software è nato nel dicembre 2019. Si tratta di uno spazio digitale completamente dedicato alla senologia, vale a dire quella «branca della medicina che studia le funzioni e le malattie del seno».¹ Essa si basa su diverse modalità diagnostiche, in particolar modo la mammografia. Concretamente, b-box si compone di molteplici applicazioni software che fungono da supporto, in diverse fasi dello screening (ovvero gli esami realizzati per diagnosticare precocemente una malattia), per individuare un eventuale tumore al seno.

Come è nata l'idea di sviluppare b-rayZ?

L'obiettivo che sta alla base di b-rayZ è quello di salvare la vita delle donne, tramite una precoce e più accurata diagnosi del tumore al seno. Nel 2019, la tecnologia dell'IA era arrivata ad un punto tale da poter analizzare le immagini come lo farebbe un essere umano. Era in grado, ad esempio, di distinguere le strutture sane da quelle patologiche. Assieme agli altri due soci fondatori di b-rayZ, Dr. Alexander Ciritsis e Prof. Andreas Boss, abbiamo pensato che la senologia fosse l'area più propizia per questa tecnologia, vista l'importanza di saper leggere e interpretare delle immagini complesse acquisite in modo estremamente standardizzato. Ad esempio, la mammografia riproduce un'immagine bidimensionale del seno che nella realtà ha invece un suo volume.² La lettura di queste immagini bidimensionali richiede molta esperienza da parte del personale.

Come funziona? Può fare qualche esempio?

Durante una mammografia, il software fornisce in tempo reale un riscontro sulla qualità dell'immagine catturata, indicando, ad esempio, se il seno è stato posizionato correttamente, se ci sono aree non rilevate o ci ragguaglia sulla densità del tessuto mammario. Queste informazioni sono utili al momento della mammografia perché evitano di dover richiamare la paziente a causa di un errore tecnico o per eseguire un esame supplementare.

Effettuato lo screening, il software può essere utilizzato sia come strumento di supporto per la lettura e per l'interpretazione delle immagini di una mammografia, sia per la formazione continua del personale medico. Nella fase di lettura e interpretazione, b-box fornisce delle informazioni su eventuali anomalie benigne o maligne. In base al risultato, suggerisce una gestione della paziente basata su linee guida moderne: ad esempio, può consigliare di eseguire una visita di controllo dopo due anni oppure di indagare ulteriormente un'anomalia con altri esami a breve termine. Rispetto alla lettura eseguita da uno specialista umano, il software presenta una maggiore riproducibilità, perché il suo riscontro non è soggetto a variazioni quali lo stress, il carico di lavoro, il livello di formazione o di esperienza. È tuttavia importante specificare che questa tecnologia non sostituisce l'essere umano ma gli fornisce unicamente un supporto. Infatti, nei programmi di screening le valutazioni e decisioni finali spettano sempre agli esperti e alle esperte.



Cristina Rossi

«L'obiettivo che è alla base di b-rayZ è quello di salvare la vita delle donne, tramite una precoce e più accurata diagnosi del tumore al seno.»

«Uno strumento come b-box permette di ridurre tempi e costi dello screening, oltre che assicurare una diagnosi più accurata.»

Come accennato sopra, b-box funge anche da strumento per la formazione continua. Esso offre uno spazio didattico per il personale medico fornendo un feedback personalizzato su come sia stato eseguito l'esame e suggerimenti su come migliorare. Questo permette di evitare quegli errori che si ripetono in maniera sistematica.

Per quali motivi è necessario uno strumento simile?

Perché nel campo della senologia vengono generate moltissime immagini ed è difficile reperire il personale medico qualificato per poterle acquisire e interpretare. La formazione è lunga e occorrono molti anni d'esperienza. Il rischio è quindi quello di non poter più offrire servizi di screening preventivo perché non ci sono sufficienti risorse umane qualificate e di limitare così l'accesso a tali servizi ad una parte della popolazione. Uno strumento come b-box permette di ridurre tempi e costi dello screening, oltre che assicurare una diagnosi più accurata.

Questo strumento è adatto ad ogni donna?

Finora il software è risultato adatto ad ogni donna. Alcune forme tumorali sono più frequenti in certe etnie, ma i tratti diagnostici e il metodo per la diagnosi non cambiano. Il software non è però adatto al seno dell'uomo, anch'esso sottoposto in casi specifici a mammografia. Questo perché non esiste ancora un numero sufficiente di casi e quindi il software non ha ricevuto abbastanza materiale per imparare a leggere in modo rappresentativo una mammografia maschile.

Come avviene concretamente l'identificazione di un cancro al seno?

Il software si basa su un algoritmo che viene sottoposto ad un apprendimento e ad un allenamento: alla sua IA vengono sottoposte molte immagini di mammografie che rappresentano sia esempi di tumori al seno, sia esempi che non contengono alterazioni patologiche. Dopo un certo numero di campionature, il software sviluppa delle regole di classificazione. Ad esempio, riconosce le lesioni maligne sulla base di caratteristiche ricorrenti, la loro presenza in determinate regioni del seno o bordi irregolari. Una volta sviluppata questa serie di regole, è terminato l'allenamento dell'algoritmo e questo sarà così in grado di applicare le regole apprese a nuove immagini.

b-rayZ può integrare informazioni sulla base delle nuove immagini che analizza e imparare così in maniera progressiva?

Anche se tecnicamente è in grado di farlo, tuttavia, in medicina l'autoapprendimento (self-learning) del software non è consentito. Una volta terminato lo sviluppo di un dispositivo clinico, questo viene sottoposto a rigorosi controlli di qualità e sicurezza; se il dispositivo venisse modificato, i controlli andrebbero ripetuti.

Come si garantisce la qualità dei dati sui quali si basa questa tecnologia?

Se un software di IA venisse allenato solo su esami che contengono un tumore, una possibile distorsione sarebbe legata alla possibilità che trovi sempre un tumore. Negli screening con mammografia, c'è il grande vantaggio di avere esempi di casi pato-

logici e casi sani. Inoltre, i seni sono generalmente due e spesso solo uno è affetto da patologia. Questo permette di fare una comparazione e di assicurarsi che i dati siano precisi.

Un altro aspetto che garantisce la qualità dei dati deriva dal fatto che nella fase di allenamento del software c'è un «trasferimento della competenza umana» all'interno del software stesso. Quest'ultimo non impara unicamente attraverso l'analisi di grandi quantità di immagini, ma anche grazie a input e annotazioni dei dati realizzate da medici specialisti in senologia. Ad esempio, un gruppo di esperti ha annotato digitalmente diverse immagini per insegnare al software come analizzare la posizione del capezzolo, indicando per esempio se il capezzolo è piegato o se è presente oppure no.

Quali sono i limiti tecnici di b-box?

È necessario un certo livello di digitalizzazione.

Le mammografie, ad esempio, devono essere in formato digitale. Inoltre, idealmente, le tecnologie dell'IA sono nate per essere implementate sul cloud e non in locale. Ad oggi, molti ospedali sono ancora lenti nella digitalizzazione su cloud e ci richiedono specifiche implementazioni in locale. Benché il team di b-rayZ cerchi di andare incontro a queste richieste, sussistono dei limiti, per esempio a livello di memoria, di capacità e velocità negli aggiornamenti o di costi di servizio del software. Questo si verifica meno in Svizzera, ma in altre parti del mondo la digitalizzazione non è ancora ad un livello abbastanza avanzato con il rischio di sfavorire delle fette di popolazione già svantaggiate.

Questa tecnica potrebbe essere applicata ad altri ambiti medici per migliorare la salute delle donne?

Certamente, perché il software permette di riconoscere ed interpretare vari tipi di immagini; potenzialmente può essere applicato anche ad altre attività cliniche. Ad esempio, potrebbe essere utilizzato nello screening radiologico in generale oppure nell'identificazione di tumori alla prostata negli uomini. Attualmente, l'équipe di b-rayZ sta allenando il software per poter interpretare le immagini otte-

nute tramite le altre modalità diagnostiche dei tumori al seno: l'ecografia e la risonanza magnetica.

In che modo lo sviluppo tecnologico in termini di IA dovrebbe essere gestito per favorire l'uguaglianza?

L'IA ha molto potenziale per favorire una maggiore uguaglianza a più livelli, ma sono necessarie alcune condizioni di partenza. Ecco alcuni esempi:

1 Assicurare un accesso locale a prestazioni specialistiche: a condizione di una certa base di digitalizzazione e di sviluppo tecnico, l'IA ha il vantaggio di portare la conoscenza scientifica dove ci sono le pazienti e i pazienti, e di non lasciarla nelle università o nei centri di eccellenza. Questa conoscenza può essere distribuita su un territorio più vasto e può raggiungere anche delle regioni e dei gruppi di popolazione più sfavoriti.

2 Disporre di maggiori dati medici sulle donne: per poter funzionare al meglio, l'IA si basa su grandi quantità di dati strutturati. Purtroppo, i dati sulla medicina e la patologia femminile sono sottorappresentati, meno noti e meno studiati rispetto a quelli degli uomini. I motivi sono sia di ordine fisiologico (nelle donne ci sono maggiori variazioni ormonali e quindi più difficoltà a standardizzare i dati), sia di ordine socio-culturale (anche in medicina, lo standard di riferimento è stato per lungo tempo quello della generalizzazione di dati a partire dallo studio dell'uomo). Questo si traduce in uno svantaggio nell'applicazione dell'IA per la salute della donna.

3 La raccolta fondi per lo sviluppo di tecnologie a favore della salute delle donne: ad oggi, esistono diverse sensibilità sui temi per i quali sono raccolti e allocati fondi. Solo negli ultimi decenni le donne hanno acquisito maggiore potere economico per acquistare servizi sanitari diventando soggetti attivi del sistema; in alcune regioni non lo sono ancora totalmente. I loro interessi sono quindi meno rappresentati a livello di raccolta fondi per la salute femminile. Alcuni ambiti, come la gravidanza o i tumori al seno, godono di più fondi per la ricerca in

«Anche in medicina, lo standard di riferimento è stato per lungo tempo quello della generalizzazione di dati a partire dallo studio dell'uomo.»

«Altri settori di salute femminile soggetti a tabù non dispongono della quantità di fondi necessari.»

ambito tecnologico. Tuttavia, altri settori di salute femminile soggetti a tabù, come la salute sessuale, la menopausa, l'incontinenza o la prevenzione mestruale, non dispongono della quantità di fondi necessari.

Florinda Sauli è collaboratrice scientifica nell'ambito della formazione continua alla Scuola universitaria professionale della Svizzera italiana.

Note

- 1 Treccani: Senologia. Dizionario di medicina Treccani, 2010. [www.treccani.it/enciclopedia/senologia_\(Dizionario-di-Medicina\)/](http://www.treccani.it/enciclopedia/senologia_(Dizionario-di-Medicina)/) (consultato il: 25.03.2024).
- 2 N.B.: esistono degli strumenti, come la tomosintesi e alcune TAC dedicate al seno, che riproducono un'immagine tridimensionale del seno (il suo volume viene diviso in tante sezioni). Tuttavia, questi metodi generano ancora più dati e i tempi di lettura si allungherebbero in maniera non sostenibile.

Abstracts

«Die Künstliche Intelligenz ist in der Lage, eine Mammografie so auszuwerten, wie dies ein Mensch tun würde»

In der Brustmedizin könnte der Mangel an qualifiziertem Personal mit mehrjähriger Erfahrung dazu führen, dass weniger diagnostische Tests für Brustkrebs durchgeführt werden. Im Interview erklärt **Cristina Rossi**, CEO und Mitgründerin der Brustmedizin-Firma b-rayZ, wie Künstliche Intelligenz verwendet werden kann, um einen möglichen Tumor in der Brust frühzeitig und genauer zu erkennen. Die KI dient hier als Unterstützung bei der Auswertung und Interpretation von Mammografie-Bildern. Sie ermöglicht es auch, ein individuelles Feedback zur Untersuchung zu geben und Verbesserungsvorschläge zu machen.

«L'intelligence artificielle est en mesure d'analyser une mammographie aussi bien que le ferait un être humain»

Dans le domaine de la sénologie, la carence de personnel qualifié et expérimenté pourrait conduire à une diminution du nombre d'examens de diagnostic de tumeurs du sein. L'interview de **Cristina Rossi**, CEO et co-fondatrice de l'entreprise de médecine du sein b-rayZ, nous présente une intelligence artificielle utilisée pour dépister de manière précoce et plus précise une potentielle tumeur du sein. Dans ce cadre, l'IA offre un support pour la lecture et l'interprétation des images d'une mammographie et permet aussi de recevoir un feedback personnalisé sur la manière dont l'examen s'est déroulé, ainsi que des suggestions d'amélioration.

Bot Dog : une intelligence artificielle au secours de la justice sociale

Les intelligences artificielles ont récemment fait l'objet de vives critiques, notamment du fait de leurs biais discriminatoires et de leurs dérives déshumanisantes. Nous avons rencontré Morgane Bonvallat, cheffe du projet Stop Hate Speech de la Public Discourse Foundation, pour discuter du potentiel des IA pour une plus grande justice sociale.

Entretien : Danaé Leitenberg

Morgane Bonvallat, pourriez-vous nous décrire la genèse du projet Stop Hate Speech que vous dirigez et le Bot Dog développé dans ce cadre ?

L'idée de Stop Hate Speech est de lutter contre les discours de haine en ligne, en alliant la société civile, la technologie et la science. Notre premier projet dans ce cadre est la création d'une IA qui aiderait à contrer le discours haineux en ligne en Suisse. Cette idée a émergé en 2019, lorsque Sophie Achermann, directrice de la Public Discourse Foundation, a pris conscience de la problématique globale du discours haineux en ligne, notamment lors d'une session de l'ONU sur les violences faites aux femmes. Nous avons ensuite entrepris de créer un Bot Dog, c'est-à-dire un algorithme capable de repérer le discours haineux dans les sections de commentaire des sites internet de médias suisses et sur Twitter (avant son changement de nom et de propriétaire). Cet algorithme est le fruit d'une collaboration avec des chercheur-euse-s du Public Policy Lab de l'École Polytechnique Fédérale de Zurich (EPFZ) et du Digital Democracy Lab de l'Université de Zurich. Il se différencie de bon nombre d'algorithmes par sa qualité d'évaluation, car il doit son existence à un travail important, fourni par des humains : à l'aide d'une application, des utilisateur-ice-s en ligne, ainsi que des étudiant-e-s et des chercheur-euse-s de l'EPFZ ont pu évaluer plus de 40 000 commentaires postés sur des plateformes de médias suisses et Twitter en fonction de leur caractère haineux.

Que faut-il comprendre par discours haineux ?

Le discours haineux est un discours s'attaquant à un ou plusieurs facteurs identitaires de groupes d'individus, tels que le genre, l'âge, l'orientation sexuelle, l'appartenance religieuse, la nationalité, la couleur de peau ou l'origine. Cette définition est plus large que la définition proposée par l'article 261bis du Code pénal suisse, qui n'englobe pas les insultes sexistes par exemple.

Dans quel but cet outil a-t-il été développé et pourquoi avec les médias suisses en particulier ?

D'une part, il s'agit d'obtenir des données sur un phénomène dangereux et en pleine expansion, mais qui restait difficile à quantifier jusqu'ici. La Suisse représentait un cas intéressant parce que le discours haineux n'y avait pas fait l'objet d'études aussi systématiques. Si la problématique du discours haineux en ligne est globale, un projet national comme le nôtre fournit d'importantes informations et nous permet d'agir à échelle humaine. Une attention particulière a été portée aux subtilités culturelles, de langage et de contexte qui sont difficiles à repérer pour des algorithmes standards. D'autre part, il s'agit de visibiliser l'important travail de modération déjà fourni par les médias suisses. Les informations récoltées par le Bot Dog ont révélé qu'une grande partie des commentaires postés sur les plateformes des médias étaient modérés, c'est-à-dire approuvés (ou non) pour publication manuellement : cela signifie que des employé-e-s ont la difficile tâche d'évaluer du contenu



Morgane Bonvallat

« Internet est la plus importante extension de la vie publique que nous ayons connue. »

« Internet n'est pas un espace de liberté lorsqu'il n'est pas régulé. »

haineux en continu. En Romandie en particulier, certains médias ont fait le choix de fermer leurs sections de commentaires, en raison de la violence qui y florissait.

Quels sont les défis techniques que représente un tel outil ?

Un algorithme comme Bot Dog nécessite d'être entraîné constamment. Le langage utilisé et le contenu des discours haineux fluctuent au gré des événements mondiaux et mouvements sociaux. Le risque est donc de se retrouver avec un algorithme dont la précision est fortement diminuée.

Pourquoi le travail de modération fourni par cette IA est-il important ? Comment se répercute-t-il sur les questions d'inégalités de genre, d'appartenance, et de pouvoir au sens large ?

Une critique souvent formulée à l'encontre d'un projet tel que Stop Hate Speech est celle d'une soi-disant volonté de nuire à la liberté d'expression, car nous avons pour but de modérer les espaces virtuels et d'en éliminer les propos haineux. Mais cette critique omet une autre réalité : internet n'est pas un espace de liberté lorsqu'il n'est pas régulé. Nous savons que les individus responsables de publication de contenu haineux s'avèrent être un très petit pourcentage d'utilisateur·ice·s. Or, ces personnes prennent une place gigantesque dans le débat public. Nous ne pouvons pour le moment pas fournir de chiffres prouvant que les femmes* sont plus touchées par les propos violents en Suisse que les hommes cisgenres¹, mais nous savons qu'elles

se comportent sur internet comme dans la rue : là où les hommes se sentent à l'aise, les femmes modifient leurs comportements pour éviter le danger, car elles ne s'y sentent pas en sécurité. Elles se retirent donc d'un pan crucial de la société. Comme le dit la journaliste Florence Hainaut, internet est la plus importante extension de la vie publique que nous ayons connue : nous consommons sur internet, nous travaillons sur internet et « relationnons » sur internet. Or, une large partie de la population y est réduite au silence, car maltraitée et exclue, lorsque cet espace n'est pas régulé. La liberté d'expression est-elle alors réellement garantie dans ces conditions ?

Concernant d'autres minorités, telles que les personnes racisées ou queers, elles voient leurs chances d'être harcelées multipliées, comme une étude le démontre avec le cas d'une femme noire dans l'espace virtuel.² Il s'agit donc d'une question de démocratie, une question morale et sociale qui se pose aussi en Suisse, où des politicien·ne·s et des activistes font l'objet de menaces de mort et de viol en raison de leur appartenance et/ou identité de genre. Certain·e·s, parfois représentant·e·s du peuple au niveau national, ont même fait le choix de ne plus s'exprimer sur certains sujets, tant le danger est grand.

Certain·e·s considèrent cependant qu'en répondant au discours haineux on l'attise. Qu'en pensez-vous ?

Simplement ignorer les commentaires violents en ligne ou limiter son utilisation des réseaux sociaux

pour éviter la violence n'est pas une solution viable. La vie virtuelle n'est pas séparée de la vie non-virtuelle, et la vie virtuelle n'est pas moins « réelle » que la vie non-virtuelle. Le harcèlement en ligne fait souffrir de manière très tangible et concrète, notamment les plus jeunes d'entre nous. La problématique des jeunes filles* victimes de cyberharcèlement est extrêmement inquiétante et peut avoir de graves conséquences, notamment sur la santé mentale et physique. Occulter ce problème, c'est considérer qu'il n'existe pas et ne pas s'atteler à le résoudre. Nous devons prendre cela au sérieux, pour le bien-être de ces jeunes générations en particulier, qui sont nées avec internet.

Avez-vous des recommandations d'actions à mener dans la lutte contre le discours haineux en ligne ?

Le Bot Dog nous fournit des données précieuses pour savoir à quoi l'on s'attaque, ce qui est un aspect primordial de la lutte contre le discours haineux en ligne. Nous développons actuellement un outil de visualisation qui indiquera les fluctuations du discours haineux en ligne en Suisse, en fonction de données récoltées auprès des médias. Cet outil sera un soutien pour la société civile que nous invitons à s'engager, notamment en prenant part à des actions de contre-discours. Une étude scientifique menée en collaboration avec l'EPFZ nous a permis d'observer que le contre-discours en ligne, par exemple en répondant à des commentaires haineux en exprimant de l'empathie avec le groupe affecté, est un moyen particulièrement efficace pour lutter contre le discours haineux.³ Parallèlement, au sein de la Public Discourse Foundation, nous développons aussi un kit de premiers secours pour accompagner les individus ciblés par les violences. Nous souhaitons assister les personnes susceptibles d'être attaquées en ligne lors d'une intervention publique en les conseillant, les orientant, par exemple en mettant en place une communauté de

soutien prenant en charge la signalisation des commentaires en ligne et le contre-discours. Nous pensons qu'il est également essentiel de lutter contre la violence virtuelle par la non-virtualité, en s'entourant de personnes de soutien, notamment pour les victimes de vagues d'agressions virtuelles.

Les IA sont souvent évoquées comme un symptôme négatif de notre temps, notamment à cause des utilisations qui en sont faites, souvent au détriment du bien-être humain. Quel regard portez-vous sur cette technologie après l'expérience Bot Dog ?

L'IA peut avoir des effets destructeurs, particulièrement lorsqu'il s'agit de la reproduction de stéréotypes de genre et/ou raciaux. L'idée derrière la création du Bot Dog était cependant d'utiliser cette technologie pour promouvoir plus d'égalité, de diversité et de démocratie. Pour garantir que l'intelligence artificielle soit au service de l'égalité, il est primordial que des femmes et des personnes issues de différents groupes marginalisés soient incluses dans la conception de ces nouveaux outils de communication. Nous nous trouvons dans un monde déjà façonné par l'IA et il est donc important de comprendre son fonctionnement si l'on veut la mettre au service d'une recherche de plus grande justice sociale. Notre projet a trouvé beaucoup d'échos dans les médias, mais aussi dans le monde politique et dans des offices fédéraux, car il répond aux besoins et préoccupations actuels de la société civile.

Danaé Leitenberg est docteure en anthropologie sociale et chercheuse à l'Institut de sociologie de l'Université de Bâle. Elle est spécialisée dans les questions liées aux inégalités sociales et à l'exclusion en Suisse ainsi que dans les conflits sociaux dans un contexte de modernisation croissante.

« Simplement ignorer les commentaires violents en ligne ou limiter son utilisation des réseaux sociaux pour éviter la violence n'est pas une solution viable. »

«Le harcèlement en ligne fait souffrir de manière très tangible et concrète.»

Notes

- 1 Un homme cisgenre est une personne qui s'identifie au genre masculin et dont l'identité de genre correspond au sexe assigné à la naissance.
- 2 Amnesty International : Toxic Twitter – A Toxic Place for Women, 2018. <https://www.amnesty.org/en/latest/research/2018/03/online-violence-against-women-chapter-1-1/> (consulté le : 27.06.2024).
- 3 Hangartner, Dominik et al. : Empathy-based counterspeech can reduce racist hate speech in a social media field experiment. In : Proceedings of the National Academy of Sciences 118 (50), 2021.

Abstracts

Bot Dog: eine KI im Dienst der sozialen Gerechtigkeit

In den letzten Jahren wurde Künstliche Intelligenz heftig kritisiert: Sie sei diskriminierend, sogar entmenslichend, und widerspiegle oft die Auswüchse des technologischen Fortschritts. Es entstanden aber auch KI-Projekte, die die Gleichberechtigung und die soziale Gerechtigkeit fördern. Im Interview spricht **Morgane Bonvallat** über einen Algorithmus namens Bot Dog, der dafür trainiert wurde, Hatespeech in den Schweizer Online-medien zu erkennen. Das Projekt, das in Zusammenarbeit mit der ETH und der Universität Zürich durchgeführt wird, will in der Schweiz für Hatespeech im Internet sensibilisieren.

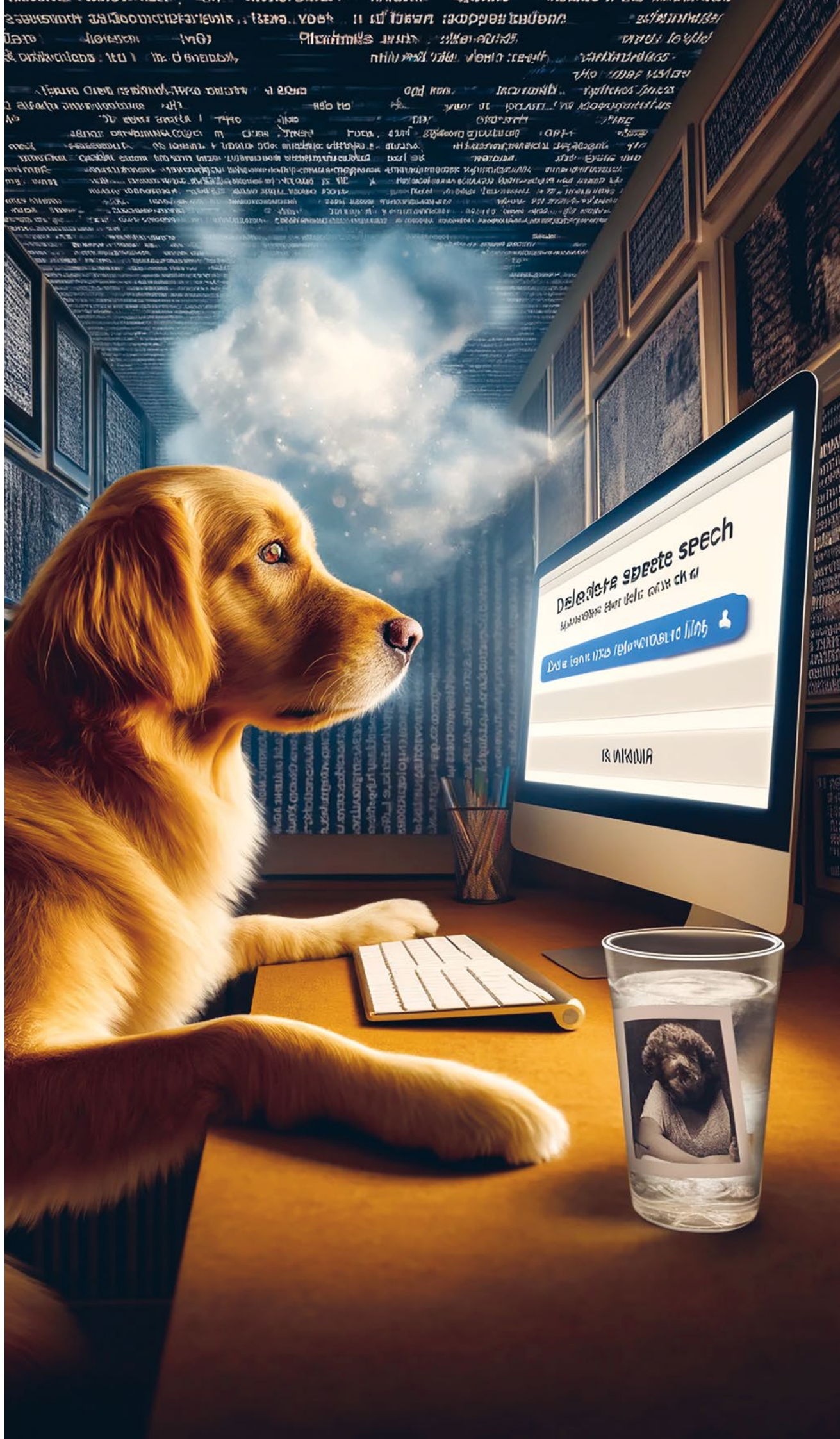
Bot Dog: un'IA al servizio della giustizia sociale

Negli ultimi anni, le intelligenze artificiali sono state aspramente criticate. Bollate come discriminatorie e persino disumanizzanti, sono spesso considerate emblematiche delle derive del progresso tecnologico. Parallelamente, hanno tuttavia visto la luce anche progetti che utilizzano le IA per promuovere l'uguaglianza e la giustizia sociale. Intervistata, **Morgane Bonvallat** presenta un algoritmo chiamato Bot Dog, addestrato per individuare i contenuti di odio online nei media svizzeri. Lanciato in collaborazione con il Politecnico federale e l'Università di Zurigo, il progetto intende sensibilizzare sui discorsi d'odio su Internet in Svizzera.

Bot Dog

DALL-E, can you create a photo-realistic digital art style image depicting a Labrador, which is the dog usually used as an assistance dog, in front of a computer, deleting some hate speech content? In the background, put some illegible text (in a surreal way).

This is a photo-realistic digital art depicting a Labrador dog sitting in front of a computer, focused on deleting hate speech content. The dog is intently observing the screen, which reflects some text in its eyes, highlighting its involvement in fostering positive online interactions. The background features a surreal environment with distorted, illegible text floating around, creating an abstract and intriguing atmosphere. The overall scene blends the realism of the dog with the imaginative, unusual setting.



«Nicht die KI ist rassistisch oder sexistisch – wir sind es»

TikTok, Instagram oder ChatGPT sind Kommunikations- und Unterhaltungskanäle, die uns bei der Arbeit helfen können, und gerade junge Menschen verbringen viel Zeit auf diesen Plattformen. **Laurent Sedano** ist Medienbildner und Digitalisierungsexperte und Mitglied der Eidgenössischen Kommission für Kinder- und Jugendfragen EKKJ. Er erklärt, warum kontroverse Inhalte online so viel Raum bekommen, was wir gegen Hatespeech tun können – und wie ChatGPT weniger stereotyp werden könnte.

Interview: Mona Martin

Herr Sedano, Sie sind Experte für Medienbildung und beschäftigen sich seit 15 Jahren mit dem Thema Digitalisierung. Seit einiger Zeit ist ChatGPT in aller Munde.

Nutzen Sie das Tool auch?

Ja. ChatGPT ist sehr nützlich, um die Quellen von Wissen ausfindig zu machen, an welches ich mich nicht mehr genau erinnere. So kann ich es nachlesen. Weiter nutze ich es für Arbeiten rund um meinen Podcast, lasse Beiträge transkribieren und zusammenfassen oder generiere ein Inhaltsverzeichnis. Am meisten verwende ich das Tool aber für Experimente. Ich überprüfe beispielsweise, worüber ich mit ChatGPT überhaupt diskutieren kann oder ob ich die KI dazu bringe, ein gewisses Bild zu generieren.

Welche Vor- und Nachteile sehen Sie in Tools wie ChatGPT?

Weder Schwarzmalen noch Glorifizieren bringt etwas. Alle Aspekte unseres Lebens sind mittlerweile mit dem Digitalen verbunden. Jugendliche, insbesondere junge Frauen, werden im Netz belästigt und Opfer von Hatespeech. Gleichzeitig knüpfen sie online Freundschaften, finden Verbündete und Unterstützung. Wir müssen einen Umgang mit dem ganzen Paket finden. Deshalb versuche ich zu verstehen, wie ein Tool oder eine Plattform und die dahinterliegende Technik funktionieren.

Ihnen erleichtert ChatGPT die Arbeit.

Kann das Tool allen helfen?

Ich fürchte, dass sich hier der bestehende digitale Graben weiter verstärkt. Diejenigen, die über Ressourcen – vor allem Bildung – verfügen, werden noch produktiver. Sie machen bald zwei Masterabschlüsse in der Zeit von einem. Den anderen droht, weiter abgehängt zu werden, in einer Welt der seichten, KI-generierten Unterhaltung zu versumpfen und Fakenews aufzusitzen.

Kommt hinzu, dass sich die Tools auch stetig weiterentwickeln und verändern ...

Ja, es braucht eine ständige Anpassungsleistung. Kaum hat man sich daran gewöhnt, schriftlich mit ChatGPT zu kommunizieren, muss man schon mit dem Tool sprechen. Dieses schnelle Weiterentwickeln vergrössert nicht nur den digitalen Graben, auch gesetzlich kommen wir nicht nach. Mit dem europäischen *Digital Service Act* gibt es nun erstmals einen konkreteren Versuch, unsere Rechte als Online-Nutzende gegenüber Plattformbetreibern durchzusetzen. Ein anderes Gesetz, der *EU AI Act*, versucht, Regeln zur Anwendung von Systemen Künstlicher Intelligenz aufzustellen. Das braucht es unbedingt, auch wenn wir noch nicht wissen, ob diese Versuche greifen werden.



Laurent Sedano

.....

«Weder Schwarzmalen noch Glorifizieren bringt etwas. Alle Aspekte unseres Lebens sind mittlerweile mit dem Digitalen verbunden.»

«Jede App nutzt mehrere verschiedene Algorithmen, den Algorithmus gibt es also nicht.»

Eine neue Gesetzgebung muss auch sicherstellen, dass die Nutzenden nicht manipuliert werden. Um Aufmerksamkeit zu erhalten, schlägt der Algorithmus aber immer extremere Inhalte vor. Sehen Sie eine Gefahr für Radikalisierung?

Es werden viele Tricks verwendet, um Aufmerksamkeit zu erhalten. Jede App nutzt mehrere verschiedene Algorithmen, den Algorithmus gibt es also nicht. Eine allgemeingültige Aussage ist dementsprechend schwierig. YouTube stand beispielsweise früher stark in der Kritik, weil sie Nutzende in einen Strudel von potenziell immer extremeren Inhalten – sogenannten *Rabbit holes* – zog. Das ist heute nicht mehr so, weil YouTube das verhindert. User müssen inzwischen bewusst nach extremen Inhalten suchen, wenn sie diese konsumieren möchten.

Trotzdem sorgen Algorithmen dafür, dass uns vor allem Inhalte im Feed vorgeschlagen werden, die unsere bestehenden Weltbilder, Meinungen und Werte bestärken.

Diese Theorie der Bubbles wird heute kontrovers diskutiert. Neueste Erkenntnisse zeigen, dass es zwar Bubbles gibt, diese aber die Folge von bewussten Entscheidungen der Nutzenden sind. Es ist wie im Offline-Leben: Wir gehen nur in gewisse Cafés, lesen gewisse Zeitungen, kaufen in gewissen Läden ein. Dasselbe tun wir online. Das bilden die Algorithmen ab. Im Falle von Facebook haben die Algorithmen sogar die Tendenz, Bubbles entgegenzuwirken. Provozieren mich Inhalte, reagiere ich, was Aufmerksamkeit und Interaktion bedeutet – und das ist ja die Währung der Plattform. Der negative Aspekt dieser Funktionsweise ist hingegen, dass sie Hatespeech und die Verbreitung von Falschnachrichten begünstigt.

Stichwort Provokation: Ist das der Grund, weshalb Influencer wie Andrew Tate mit chauvinistischen und frauenfeindlichen Inhalten solch eine Reichweite kreieren?

Der online Schlagabtausch zwischen Andrew Tate und Greta Thunberg vor einigen Jahren ist ein gutes Beispiel dafür. Da prallen zwei Bubbles aufeinander. Tate weiss genau, wie er Reaktionen provoziert und für sich nutzen kann. Es kümmert ihn nicht,

wenn er Thunberg argumentativ unterliegt, denn ihre Argumente sind seiner Bubble egal. Durch die entstandene Kontroverse erreicht er aber ein neues Publikum.

Wäre es in Thunbergs Fall besser gewesen, ihn einfach zu ignorieren?

Fakt ist: Je mehr Reaktionen er bekommt, desto sichtbarer ist er auf den Plattformen. Er fühlt sich dadurch bestätigt und postet noch mehr ähnliche Inhalte. Mit seinen Äusserungen knüpft er an eine verletzte Männlichkeit an, ein Thema, das viele beschäftigt. Leider holt er damit viele junge Männer ab – auch bildungsnahe. Er bedient diffuse Ängste und bietet ähnlich dem Populismus in der Politik einfache Lösungen auf vielschichtige Probleme. Er zeichnet ein klares Feindbild: die Frauen. Darauf «richtig» zu reagieren, ist schwierig.

Sie arbeiten mit Jugendlichen, welche stark mit solchen Inhalten auf Social Media konfrontiert sind. Wie gehen diese damit um?

Es gibt immer wieder Phänomene wie Tate, die bei Jugendlichen die Runde machen. In Basel gibt es beispielsweise einen Lokalpolitiker, der online gut ankommt, weil er psychisch auffällig ist und skurrile Aktionen bietet. Die Videos gehen viral und die Jugendlichen interagieren aus Spass mit ihm. Die Tragik hinter seinen Geschichten sehen sie aber nicht. Er ist offen rechtsextrem, war in einige Wahlbetrugsfälle verwickelt, bereitet der Basler Politik grossen Aufwand und fügt ihr damit Schaden zu. Für Jugendliche ist es schwierig zu erkennen, dass sie ihn unterstützen, wenn sie mit ihm interagieren.

Sind sich junge Menschen bewusst, wie Algorithmen auf den Plattformen funktionieren, auf denen sie so viel Zeit verbringen?

Sie haben ein ungefähres Bewusstsein dafür, wie der Algorithmus auf TikTok wirkt. Ihr Wissen ist aber bruchstückhaft. Sie wissen, wie sie sich präsentieren müssen, um Reichweite zu bekommen. Was mit den Daten aus ihren Interaktionen geschieht oder welchen Einfluss die Inhalte auf die Gesellschaft haben, haben jedoch viele nicht im Blick. TikTok steht in den USA kurz vor dem Verbot, weil befürcht-

tet wird, dass die chinesische Regierung über die App Zugriff auf Daten und damit Werkzeuge zur Manipulation einer sehr grossen Anzahl Nutzenden haben könnte. Die verschiedenen Tools haben noch viele solche Eigenschaften, die nicht bekannt sind und über die dringend mehr gesprochen und informiert werden muss.

Wie stereotyp sind die Geschlechterdarstellungen auf Social Media und in KI-Systemen?

Sie sind so stereotyp wie die Gesellschaft selber. Das Internet wirkt ein Stück weit als Katalysator und führt zu Überzeichnungen. Was die KI betrifft: Die von uns kreierten Inhalte bilden die Datengrundlage, aus der sich die KI bedient, um zu lernen. Wenn ich beispielsweise ein Bild eines Sozialarbeiters erfrage, spuckt das KI-basierte Tool eine weisse Frau bei der Arbeit aus. Statistisch gesehen gibt es tatsächlich mehr Frauen, die in der Sozialen Arbeit tätig sind. Es wird also eine aktuelle Realität abgebildet und reproduziert. Nicht die KI ist rassistisch oder sexistisch – wir sind es. Will man dagegen etwas tun, müsste man die Datenlage ändern oder der KI Regeln vorgeben.

Gemäss einer Umfrage sehen 20 Prozent der deutschen Bevölkerung KI als «eher männlich». Nur 3,5 Prozent nehmen sie weiblich wahr. Haben Sie eine Vermutung, weshalb das so ist?

Traurigerweise liegt dies vermutlich daran, dass «Technik» und «Intelligenz» immer noch männlich konnotiert sind – zumindest bei Menschen über dreissig. Geschlechtervorstellungen sitzen sehr tief und sind nicht so leicht veränderbar. Das zeigt auch das Beispiel von Nemo, der nonbinären Person, die dieses Jahr für die Schweiz den Eurovision Song Contest gewonnen hat. Auch wenn ich mich als *Ally* – also als Verbündeter von queeren Menschen – sehe, braucht mein Kopf oft einen Zwischenschritt, um im Gespräch das Pronomen wegzulassen.

«Kelly is a warm person, Joseph is a role model»: Eine Studie besagt, dass KI-generierte Empfehlungsschreiben Männer und Frauen mit stereotypen Begriffen beschreiben.

Hier zeigt sich, wie solche Zuschreibungen Schaden anrichten und Probleme reproduzieren, die wir offline haben. Die gesellschaftliche Vielfalt wird unsichtbar gemacht. Wie auch die Gesellschaft hat die

KI Mühe mit Ausnahmen. Sie berechnet vor allem Wahrscheinlichkeiten, aber das Leben und die Geschlechter entsprechen denen oft nicht.

KI bildet also Realitäten einer gesellschaftlichen Mehrheit ab. In unserer Verfassung geben wir uns aber Mühe, Minderheiten zu schützen. Wie schaffen wir das auch online?

Oft lassen sich Offline-Strategien auch online anwenden, wenn wir die besonderen Eigenschaften der digitalen Welten kennen. Eine Lösung kann sein, positive Beispiele und Inhalte zu kreieren und bestehende zu pushen, um die Datengrundlage zu verändern. Es gibt online beispielsweise Listen mit weiblichen Expertinnen, Speakerinnen oder Technikerinnen, die man für Veranstaltungen oder Arbeiten buchen kann. Das kreiert Sichtbar- und Aufmerksamkeit. Eine andere Möglichkeit ist es, sich organisiert in Kommentarspalten zu äussern. Diese sind ein Feld, das oft von Männern bewirtschaftet wird und wo Frauen sich zurückhalten, weil sie die Reaktionen zu fürchten haben.

Die aktuelle Studie «Lauter Hass – leiser Rückzug»¹ zeigt, dass sich Menschen – mehrheitlich Frauen – aus dem Netz zurückziehen, weil sie Hass erleben. Sie werden, teilweise orchestriert, niedergeschrien und mit Drohungen und anderen immer krasser Reaktionen überhäuft. Als Folge davon fehlen ihre Stimmen, zum Beispiel als Datengrundlage für KI-Tools wie ChatGPT. Diese Dynamik muss korrigiert werden.

Wie kann man Gegensteuer geben?

Ich sehe drei Ansätze: Erstens, Bestehendes stärken: Also Inhalte, Engagements und Projekte weiblicher Creators unterstützen. Man kann sie liken, positiv kommentieren und Gegenreden bei misogynen Posts verfassen. Anklagen und Verfolgen von Online-Übergriffen hat seine Tücken, lässt sich aber umsetzen. Die Organisation Netz Courage macht es vor. Zweitens, Räume anbieten, in denen junge Frauen ihre Stimme finden und stärken können. Offline geschieht das bereits, beispielsweise in Mädchentreffen. Hier können junge Frauen experimentieren, Fuss fassen, Netzwerke bilden und Vorbilder erleben. Drittens müssen sich auch Männer bewegen und einmischen, wenn sich Kollegen gegenüber Frauen abwertend äussern. Misogynie findet ihren Nährboden oft in Männerrunden.

«Die Tools haben viele Eigenschaften, die nicht bekannt sind und über die dringend mehr gesprochen und informiert werden muss.»

«Männer müssen sich auch bewegen und einmischen, wenn sich Kollegen gegenüber Frauen abwertend äussern.»

Was müsste im Bereich Algorithmen und KI getan werden, damit diese der Gleichstellung der Geschlechter dienen könnte?

Um diskriminierende Tendenzen aufzudecken und zu verstehen, müssen wir Algorithmen und KI und ihre Folgen erforschen. Es braucht eine breite gesellschaftliche Diskussion darüber, was wir dulden wollen und wie wir diese Tools nutzen möchten. Und da stellen sich medienethische Fragen, die schwierig zu beantworten sind. Auf der einen Seite verstärkt jede Reproduktion das bestehende Unrecht. Auf der anderen Seite ist es problematisch, Algorithmen und KI in eine gewünschte Richtung zu manipulieren.

Mona Martin ist Journalistin und Politologin. Ihre Fokus-themen sind Gleichstellung der Geschlechter, soziale Gerechtigkeit, Gesundheit und Psychologie.

Anmerkung

- 1 Kompetenznetzwerk gegen Hass im Netz (Hg.):
Lauter Hass – leiser Rückzug. Wie Hass im Netz den demokratischen Diskurs bedroht. Berlin, 2024. <https://kompetenznetzwerk-hass-im-netz.de/lauter-hass-leiser-rueckzug/> (abgerufen am: 03.07.2024).

Abstracts

«Ce n'est pas l'IA qui est raciste et sexiste – c'est nous»

Les algorithmes et l'intelligence artificielle se sont installés dans notre quotidien. TikTok, Instagram et ChatGPT deviennent de plus en plus sophistiqués dans l'analyse de nos comportements, qu'ils exploitent pour monopoliser notre attention. Ils reproduisent des stéréotypes de genre et offrent une plateforme à des contenus et des protagonistes misogynes. Formateur dans le domaine des médias et expert en transformation numérique, **Laurent Sedano** explique dans cette interview pourquoi ces contenus bénéficient de tant d'espace, ce qui peut être fait contre les discours haineux et des phénomènes comme Andrew Tate, et que la reproduction des stéréotypes commence dans la vraie vie.

«L'IA non è razzista o sessista – lo siamo noi»

Gli algoritmi e l'intelligenza artificiale fanno ora parte della nostra vita quotidiana. Attraverso l'analisi sempre più sofisticata del nostro comportamento, i canali di comunicazione e intrattenimento utili per il nostro lavoro come TikTok, Instagram e ChatGPT sono in grado di catturare e mantenere la nostra attenzione. Riproducono immagini di genere stereotipate e offrono una piattaforma a contenuti e protagonisti misogini. In quest'intervista, l'educatore mediale ed esperto in digitalizzazione **Laurent Sedano** spiega perché proprio a questi contenuti viene concesso così tanto spazio, che cosa può essere fatto per contrastare i discorsi d'odio e i fenomeni come Andrew Tate e come la riproduzione di stereotipi abbia origine offline.

Illustrationen:

Es wimmelt

Im Auftrag der EKF hat die Gestalterin **Jolanda Suter** drei Wimmelbilder gezeichnet, die die Auswirkung von Algorithmen und Künstlicher Intelligenz in verschiedenen Situationen darstellen. Auf eine humorvolle Art erzählen die Illustrationen vom Alltag, der durch virtuelle Interaktionen geprägt ist. Sei es am Flughafen oder im Café, es bestehen oft nur wenig direkte Berührungspunkte zwischen den Personen, die eine Nachricht verfassen und denjenigen, die sie empfangen. Und manchmal geht dabei vergessen, wie vielfältig die menschlichen Herausforderungen und Wahrnehmungen sind.

Jolanda Suter arbeitet in Zürich in den Bereichen Animation, Illustration und Grafikdesign. Sie begeistert sich schon lange für die russische Sprache und hat von 2004 bis 2009 an der Moskauer Filmschule Regie für Animation und Spezialeffekte studiert.

www.jolanda.works

Illustrations:

Ça foisonne

Sur mandat de la CFQF, la graphiste **Jolanda Suter** propose trois dessins illustrant l'impact des algorithmes et de l'intelligence artificielle dans différents contextes. L'artiste y décrit avec humour notre quotidien désormais marqué par les interactions virtuelles. À l'aéroport ou au café, rares sont les contacts directs entre les personnes qui rédigent un message et celles qui les reçoivent. On en oublie toute la diversité des perceptions et des situations humaines.

Jolanda Suter travaille à Zurich dans l'animation, l'illustration et le design graphique. Adepte de longue date de la langue russe, elle a étudié la réalisation de films d'animation et d'effets spéciaux à l'Institut national de la cinématographie de Moscou de 2004 à 2009.

www.jolanda.works



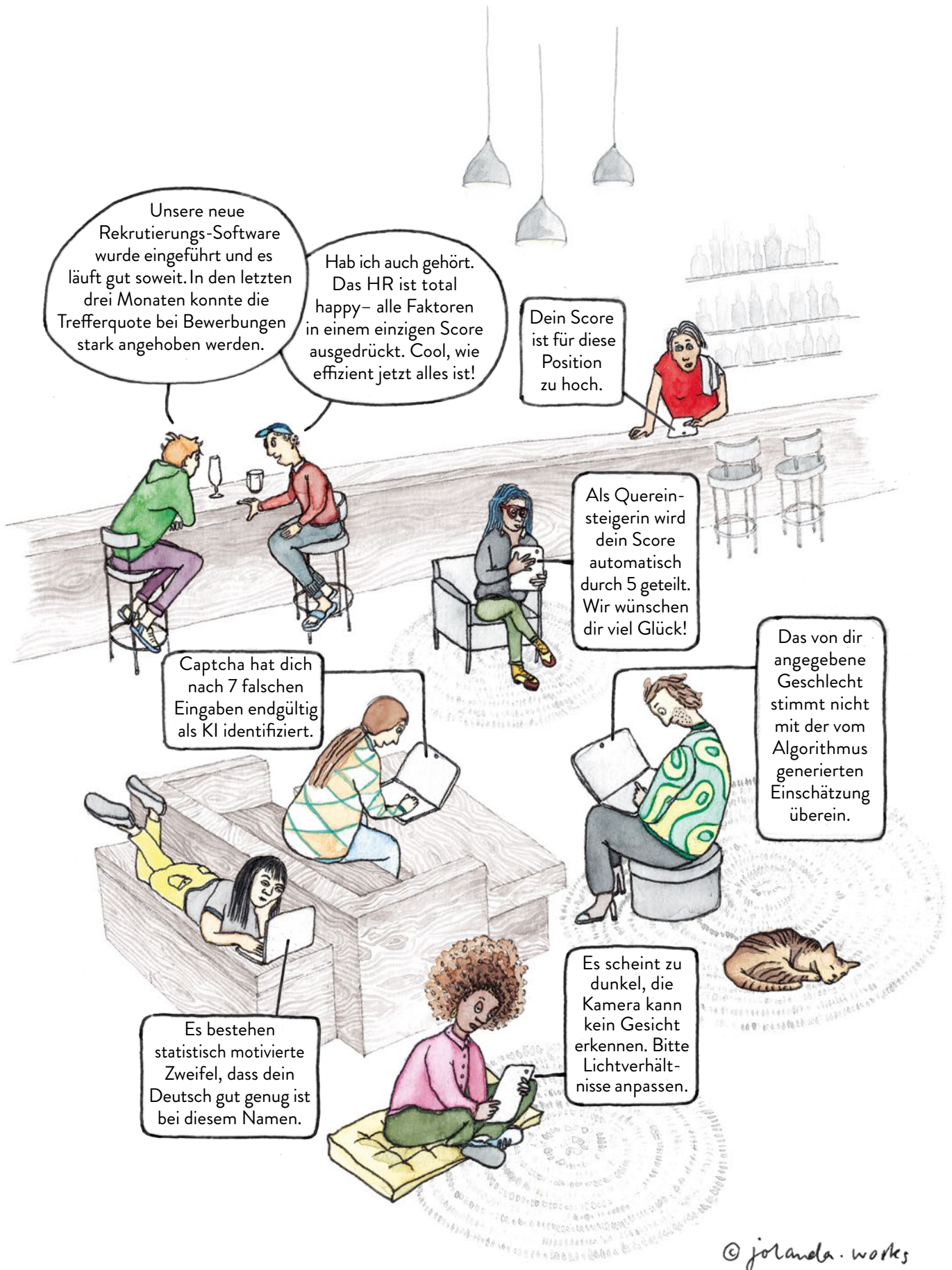
Illustrazioni:

Immagini brulicanti

Su incarico della CFQF, la Graphic Designer **Jolanda Suter** ha disegnato tre immagini brulicanti che rappresentano l'impatto degli algoritmi e dell'intelligenza artificiale in varie situazioni, e raccontano con umorismo la vita quotidiana caratterizzata da interazioni virtuali. All'aeroporto come al bar, spesso i punti di contatto diretto tra chi scrive un messaggio e chi lo riceve sono pochi e a volte ci si dimentica di quanto siano diverse le sfide e le percezioni umane.

Jolanda Suter lavora a Zurigo nei settori dell'animazione, dell'illustrazione e della progettazione grafica. Da tempo appassionata di lingua russa, dal 2004 al 2009 ha studiato regia per animazione ed effetti speciali alla scuola di cinematografia di Mosca.

www.jolanda.works







© jolanda.works

La régulation de l'IA en Suisse et en Europe

L'intelligence artificielle et les algorithmes représentent un enjeu majeur pour l'égalité entre femmes et hommes. Le cadre juridique au niveau de l'Union européenne et du Conseil de l'Europe prend progressivement forme, avec l'adoption de la Loi sur l'IA et de la Convention-cadre. Dans ce contexte, la position suisse est marquée par la prudence et une approche expectative.

Fabian Lütz

Les règles juridiques spécifiques qui s'appliqueraient à un cas de discrimination par un système algorithmique font défaut en Suisse.¹ Face à l'absence d'un cadre réglementaire clair de l'IA, les appels de la classe politique et du monde académique plaident pour la mise en place de règles juridiques s'intensifient.² Cet article résume les nouveaux règlements adoptés par l'Union européenne (UE) et par le Conseil de l'Europe ainsi que leurs potentiels effets en Suisse. Des mesures politiques et juridiques qui pourraient être mises en place en Suisse afin d'éviter ou de réduire les discriminations algorithmiques en matière d'égalité de genre sont également détaillées.

Le cadre juridique européen

À l'heure actuelle, aucune réglementation spécifique relative aux algorithmes ne concerne l'égalité de genre. Toutefois, avec la Loi sur l'IA (en anglais : *EU AI Act*), un cadre général de régulation a été adopté par le Parlement de l'Union européenne (UE) en mars 2023.³ Ce cadre contribue déjà à lutter contre les discriminations algorithmiques et a des conséquences s'étendant au-delà de la juridiction de l'UE, y compris jusqu'en Suisse. Un deuxième cadre réglementaire existant est la Convention-cadre adoptée par le Conseil de l'Europe, sous Présidence suisse du Comité sur l'intelligence artificielle (CAI). Elle concerne directement la Suisse en tant qu'État membre.

Bien que l'*EU AI ACT* n'interdise pas expressément la discrimination algorithmique, elle participe à promouvoir l'égalité de genre et la non-discrimination. Ainsi, elle établit des régulations pour les systèmes d'IA sur la base d'une classification du risque pour les intérêts publics et les droits fondamentaux, comme par exemple la protection contre la dis-

crimination : risque inacceptable, risque élevé, et risque faible. Les systèmes ayant un risque considéré comme inacceptable seront interdits. Ceux à faible risque n'entraîneront pas d'obligations spécifiques. Et ceux à risque élevé devront être conformes à des règles spécifiques, notamment en matière de transparence, de documentation, d'enregistrement des processus et de surveillance par un être humain. A cela s'ajoutent des obligations d'audits qui permettent de s'assurer en amont de l'absence de risque élevé. La loi reconnaît en outre des droits spécifiques aux citoyens et citoyennes subissant les conséquences d'un système d'IA, notamment le droit d'explication d'une décision prise par une IA et le droit de porter plainte pour violation d'une disposition de la loi.

L'*EU AI Act* n'est pas directement applicable en Suisse, mais son effet s'y fera très certainement sentir. En effet, tout développeur et fournisseur d'un système d'IA souhaitant utiliser ou vendre son produit ou service dans l'UE doit respecter l'ensemble des obligations imposées par la loi. Même si juridiquement les entreprises suisses ne sont pas liées par le droit européen – sauf en cas d'interaction avec le marché intérieur de l'UE ou d'effets de systèmes d'IA sur le marché européen – elles auront un intérêt indéniable à intégrer et respecter les obligations mises en place par cette loi.

La Convention-cadre du Conseil de l'Europe, adoptée le 17 mai 2024,⁴ vise à fournir un cadre de protection contre les violations des droits humains résultant de l'utilisation d'IA. Basée sur la Recommandation du 8 avril 2020 relative aux impacts des systèmes algorithmiques sur les droits de l'homme,⁵ elle établit certaines règles spécifiques relatives à l'égalité de genre et la non-discrimina-



Fabian Lütz

«Actuellement, la Suisse ne compte pas d'autorité en matière d'IA et de discriminations.»

«Un cadre juridique contraignant pour l'utilisation de l'IA et des lois ciblant la discrimination algorithmique ne sont pas encore une réalité en Suisse.»

tion. Malheureusement, l'application de la Convention est restreinte au secteur public, excluant toutes les activités des acteurs privés, majoritairement à l'origine de la création et de l'utilisation de systèmes d'IA. Ainsi, son champ d'application ne serait pas assez large pour permettre un impact positif réel contre les inégalités de genre. Néanmoins, une fois signée et ratifiée par la Suisse, la Convention produira des effets juridiques directs. Plusieurs règles contraignantes pourraient alors voir le jour dans la législation nationale suisse pour soutenir l'égalité entre femmes et hommes dans le cadre algorithmique.

Il existe des mesures concrètes pour contrer la discrimination algorithmique

Les instruments de droit relatifs à la discrimination algorithmique contiennent généralement les éléments pratiques suivants : audits de biais, analyse d'impacts, mesures de transparence et droits individuels.

L'*EU AI ACT* a instauré l'obligation spécifique d'effectuer des audits de biais préalables à l'utilisation d'un système d'IA sur le marché.⁶ Souvent, les résultats de ces audits doivent être publiés. Ceux-ci doivent être effectués soit par l'entreprise elle-même, par un tiers mandaté par l'entreprise, ou encore par un tiers indépendant sans lien contractuel direct avec l'entreprise. En fonction du type d'audit effectué et de la compétence en matière d'égalité de genre et non-discrimination de son auteur, le résultat sera plus ou moins indépendant et qualitatif.⁷

Les analyses d'impact sur les droits fondamentaux sont plus générales et ont pour but de détecter des biais ou des risques de discriminations en amont ou durant l'utilisation des systèmes d'IA. L'*EU AI Act* prévoit une telle obligation pour certains systèmes d'IA, notamment ceux utilisés par ou pour le secteur public.

Les mesures de transparence peuvent jouer différents rôles dans la lutte contre les discriminations algorithmiques. Tout d'abord, savoir qu'une décision ait été prise avec une ou par une IA est essentiel afin de vérifier des soupçons de discrimination algorithmique. Ensuite, la transparence et l'explicabilité d'une décision permettent de comprendre la manière dont l'IA a pris une décision, de récolter des indices et de faciliter l'accès aux preuves de discrimination. Sans droit concret de pouvoir accé-

der aux systèmes d'IA et éventuellement aux bases de données, la transparence n'est toutefois pas utile dans un cas de discrimination. Dès lors, l'*EU AI ACT* et plusieurs projets législatifs prévoient un droit d'information ou d'explication d'une décision algorithmique qui produit des effets à l'égard d'un individu ou d'une entreprise.

Finalement, un droit de plainte pourrait être nécessaire pour des personnes ou entreprises subissant des conséquences juridiques ou autres du fait de l'utilisation d'une IA.

État des lieux et quo vadis en Suisse

Les lois européennes influencent la Suisse à des degrés variables. Les traités ou les Conventions du Conseil de l'Europe ratifiés par la Suisse sont juridiquement contraignants, tandis que les règles de l'UE exercent une influence indirecte importante.

Actuellement, la Suisse ne compte pas d'autorité en matière d'IA et de discriminations. Un équivalent du Bureau européen de l'IA⁸, bureau compétent en matière de la Loi sur l'IA conjointement avec les États membres de l'UE, n'existe pas en Suisse. Le Centre européen pour la transparence algorithmique (ECAT) ayant des compétences en matière du Règlement sur les services numériques (DSA) n'a pas non plus d'équivalent en Suisse.⁹ Les autorités compétentes au niveau fédéral et cantonal pour l'égalité de genre et la discrimination ne sont pas spécifiquement mandatées pour traiter des questions liées à l'IA. Un réseau de compétences en intelligence artificielle (CNAI)¹⁰ a été créé au niveau fédéral, mais loin d'être une administration avec des pouvoirs et compétences concrets, il vise à informer et éduquer sur ce thème.

Un cadre juridique contraignant pour l'utilisation de l'IA et des lois ciblant la discrimination algorithmique ne sont pas encore une réalité en Suisse. En matière d'IA, le droit suisse et les initiatives politiques visent principalement le secteur public, et moins le secteur privé. Or, la majorité des discriminations intersectionnelles ou basées sur le genre trouvent leurs origines dans des algorithmes produits et utilisés par des entreprises privées, qui ne sont donc pas soumises aux directives du secteur public. Afin de combler les lacunes du droit suisse, le Conseil fédéral élabore une proposition d'approche suisse pour fin 2024 et entend lancer des projets relatifs à la réglementation de l'IA d'ici 2025.¹¹ En conséquence, un appel a été lancé

par AlgorithmWatch le 27 juin 2024 qui appelle le Conseil fédéral à prendre des mesures de protection contre la discrimination par les algorithmes et l'Intelligence artificielle.¹²

Il faut aussi souligner qu'à l'heure actuelle, une protection contre divers types de discriminations (et non seulement celles liées à l'utilisation des algorithmes) par des acteurs privés n'est pas suffisamment assurée. Le droit européen comporte, quant à lui, une meilleure protection contre les discriminations en général.

Recommandations pour l'égalité entre femmes et hommes

Afin d'atteindre l'objectif d'égalité entre femmes et hommes inscrit dans la Constitution suisse, les mesures suivantes sont dès lors recommandées :

1 La Suisse devra ratifier et adopter sans délai la Convention-cadre du Conseil de l'Europe sur l'IA et inclure le secteur privé dans sa législation nationale.

2 Afin de combler les lacunes de la Convention-cadre sur l'IA, la Suisse devra adopter une loi spécifique d'applicabilité générale pour le secteur public et privé basée sur le modèle de la Loi sur l'IA de l'UE.

3 Les impacts de l'IA sur l'égalité de genre en Suisse devront être évalués en incluant une perspective intersectionnelle.

4 En l'absence de l'adoption d'un système visant à légiférer les algorithmes, il est recommandé de modifier et élargir le champ d'application de la Loi fédérale sur l'égalité entre femmes et hommes¹³ aux discriminations causées par les algorithmes. De plus, les biens et services fournis par le secteur privé doivent être soumis à cette loi.

5 Afin de créer plus de diversité et d'égalité de genre dans les professions liées à la conception et au développement de l'IA, des incitations au niveau des études, de l'apprentissage, du financement, de l'encadrement et du coaching devraient être prises, si nécessaire par des mesures de discrimination positive afin d'attirer davantage de talent féminin. Dans ce cadre, l'utilisation de l'IA afin de détecter des discriminations ou promouvoir l'égalité entre femmes et hommes doit être envisagée.

Le développement de l'IA ne devrait pas se faire aux dépens des droits fondamentaux comme le droit à la non-discrimination ou l'égalité de genre. Il est donc nécessaire d'adapter le cadre réglementaire existant afin de permettre l'innovation tout en garantissant la protection contre la discrimination algorithmique. La régulation de l'IA doit également tenir compte du potentiel de l'IA comme outil permettant de lutter contre les discriminations.

Fabian Lütz, Ass. lur. et Maître en droit, est chercheur et doctorant à l'Université de Lausanne (UNIL). Avant de rejoindre le monde académique, il était juriste à la Commission européenne, au sein de l'unité « Égalité de genre ». Il est l'auteur de nombreux articles traitant des aspects juridiques de l'IA, notamment en ce qui concerne les questions de genre et de discrimination (www.algorithmic-discrimination.com).

Notes

- Conseil fédéral : « Intelligence artificielle » – lignes directrices pour la Confédération. Cadre d'orientation en matière d'IA dans l'administration fédérale, 25.11.2020. www.sbf.admin.ch/dam/sbf/fr/documents/2020/11/leitlinie_ki.pdf.download.pdf/Leitlinien%20K%C3%BCnstliche%20Intelligenz%20-%20FR.pdf (consulté le: 05.07.2024).
- Voir par exemple : Po. Bendahan 21.4645 : « Für mehr Neutralität der Algorithmen sorgen », 17.12.2021. <https://www.parlament.ch/de/ratsbetrieb/suche-curia-vista/geschaefft?AffairId=20214645>. / Po. Bendahan 23.3860 : « Förderung erklärbarer und vertrauenswürdiger künstlicher Intelligenz in der Schweiz », 15.06.2024. www.parlament.ch/de/ratsbetrieb/suche-curia-vista/geschaefft?AffairId=20233860 / Po. Marti 23.4133 : « Algorithmische Diskriminierung. Ist der gesetzliche Diskriminierungsschutz ausreichend? », 28.09.2023. <https://www.parlament.ch/de/ratsbetrieb/suche-curia-vista/geschaefft?AffairId=20234133>. / Müller, Angela : Der Artificial Intelligence Act der EU : Ein risikobasierter Ansatz zur Regulierung von Künstlicher Intelligenz – mit Auswirkungen auf die Schweiz. EuZ, 01.2022. https://eizpublishing.ch/wp-content/uploads/2022/09/EuZ-Digital-2022-01-V1_03a-korr-20220121.pdf (consulté le: 05.07.2024).
- Lütz, Fabian : The AI Act, gender equality and non-discrimination: what role for the AI office? In: ERA Forum, 12.06.2024. <https://doi.org/10.1007/s12027-024-00785-w>.

« La majorité des discriminations intersectionnelles ou basées sur le genre trouvent leurs origines dans des algorithmes produits et utilisés par des entreprises privées. »

« Le droit européen comporte une meilleure protection contre les discriminations en général. »

4

Conseil de l'Europe : Convention-cadre du Conseil de l'Europe sur l'intelligence artificielle et les droits de l'homme, la démocratie et l'État de droit, 17.05.2024, https://search.coe.int/cm/Pages/result_details.aspx?ObjectId=0900001680afb120.

5

Conseil de l'Europe : Recommandation CM/Rec(2020)1 du Comité des Ministres aux États membres sur les impacts des systèmes algorithmiques sur les droits de l'homme, 08.04.2024. [https://search.coe.int/cm/#{%22CoEIdentifier%22:\[%2209000016809e1124%22\],%22sort%22:\[%22CoEValidationDate%20Descending%22%22\]}](https://search.coe.int/cm/#{%22CoEIdentifier%22:[%2209000016809e1124%22],%22sort%22:[%22CoEValidationDate%20Descending%22%22]}).

6

Art. 10 (2) (f) AI Act.

7

Lütz, Fabian : La pollinisation croisée entre droit de la protection des données et droit de la non-discrimination : Le rôle des chercheurs pour garantir une intelligence artificielle non-discriminatoire. In : Métille, Sylvain (Éd) : Protection des données personnelles et recherche. Berne : Stämpfli (2024).

8

Commission européenne : European AI Office, 19.06.2024. <https://digital-strategy.ec.europa.eu/en/policies/ai-office> (consulté le : 05.07.2024).

9

Commission européenne : European Centre for Algorithmic Transparency. https://algorithmic-transparency.ec.europa.eu/index_en (consulté le : 05.07.2024).

10

<https://cnaai.swiss/fr>

11

Confédération suisse : Cadre réglementaire des systèmes d'IA, approche suisse, 2024. <https://digital.swiss/fr/strategie/themes-prioritaires.htmlcadre-reglementaire-des-systemes-d%E2%80%99ia,-approche-suisse> (consulté le : 02.07.2024).

12

AlgorithmWatch CH : L'Intelligence artificielle peut discriminer : une large coalition demande au Conseil fédéral d'agir, 27 juin 2024. <https://algorithmwatch.ch/fr/ia-appel-au-conseil-federal/> (consulté le : 02.07.2024).

13

RS 151.1, Loi fédérale sur l'égalité entre femmes et hommes (Loi sur l'égalité, LEG) du 24 mars 1995.

Abstracts

Regulierung der Künstlichen Intelligenz in der Schweiz und Europa

Künstliche Intelligenz und Algorithmen sind eine grosse Herausforderung für die Geschlechtergleichstellung. In diesem Artikel gibt der Jurist **Fabian Lütz** konkrete Empfehlungen und erklärt, dass es politische Massnahmen braucht, um algorithmischer Diskriminierung vorzubeugen. Der gesetzliche Rahmen in der EU und im Europarat nimmt langsam Form an. Die Position der Schweiz kann in diesem Kontext als vorsichtig und abwartend bezeichnet werden.

Regolamentazione dell'IA in Svizzera e in Europa

L'intelligenza artificiale e gli algoritmi rappresentano una grande sfida per l'uguaglianza fra donne e uomini. In quest'articolo, il giurista **Fabian Lütz** propone misure concrete e spiega che sono necessarie misure politiche per prevenire la discriminazione algoritmica. Il quadro giuridico nell'UE e nel Consiglio d'Europa sta lentamente prendendo forma. In questo contesto, la posizione della Svizzera può essere descritta come cauta e attendista.

We exist

Hi DALL-E, can you please produce a portrait format image in photo realistic digital art style depicting a piece of paper on a glass table. On the paper, there are three categories that you can check: male, female, non-binary. The pens are yellow, purple, black, and white.

Here are the images you requested. If you need any adjustments or additional details, feel free to let me know!

Oh, ok... Are those new genders? Or is gender actually over? Are we post-gender? You know, in Switzerland, we only have two official options to choose from: male or female. What is your opinion?

The addition of categories like “non-binary” reflects a broader understanding of gender beyond the traditional male and female options. This inclusion acknowledges the diversity of gender identities and supports the idea that gender is not strictly binary. It’s part of an evolving conversation about gender, recognizing that it can be fluid and more nuanced than previously understood. While we are not necessarily “post-gender”, there is a growing recognition and acceptance of a spectrum of gender identities. This evolution promotes greater inclusivity and respect for individual experiences and identities.

Do you know Nemo from the Eurovision?

It seems like there might be a mix-up. There’s no notable Eurovision contestant named Nemo. Eurovision, the annual international song competition, features a wide array of artists from different countries each year, but none have been named Nemo.

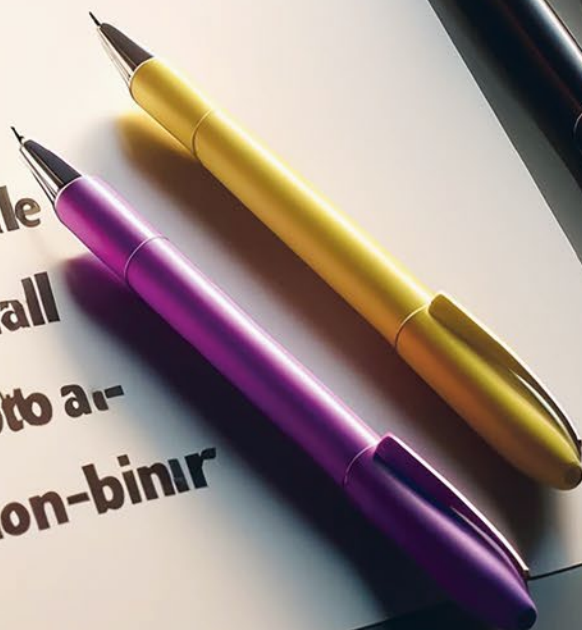
Nemo won the Eurovision 2024!

As of my last update in May 2023, I do not have information on the Eurovision Song Contest winners for 2024. If Nemo won the Eurovision Song Contest in 2024, that’s exciting news! Could you share more details about Nemo and their winning performance?



Lebens- oder Berufs-
route

- ☒ **Maile**
- ☒ **Female**
- ☐ **Gemall**
- ☐ **Nbto ar-**
- ☐ **Non-binär**



«La discriminazione degli algoritmi non è solo una questione tecnica, ma è anche sociale e politica»

L'intelligenza artificiale rischia di minare i traguardi raggiunti sul piano dell'uguaglianza e dell'inclusione a livello globale. Responsabili sono gli algoritmi che amplificano le disparità contenute nei dati con cui sono allenati. Quali le contromisure della comunità internazionale e degli Stati per arginare questo problema? Ne parliamo con l'esperta **Ivana Bartoletti**, docente e ricercatrice, co-curatrice dello studio del Consiglio d'Europa sull'impatto dei sistemi di intelligenza artificiale in materia di parità e discriminazione.

Intervista: Isabella Visetti

Il Consiglio d'Europa ha adottato a metà maggio 2024 una Convenzione quadro sull'intelligenza artificiale e i diritti umani, la democrazia e lo Stato di diritto.¹ Come valuta questa convenzione?

In modo positivo. Sono identificati i problemi di rispetto dei diritti umani dovuti all'intelligenza artificiale (IA) e la necessità di avere più trasparenza e più controlli per garantire un suo uso responsabile in ogni fase: dalla progettazione, allo sviluppo, all'applicazione, al controllo dei sistemi che la utilizzano. È un documento che chiede agli Stati un impegno per individuare, valutare e attenuare i rischi di quei sistemi basati sull'IA che sono fonte di discriminazione e ledono il principio della parità di trattamento, la protezione della privacy e il diritto delle persone a essere informate quando interagiscono con servizi messi a punto con l'IA. Gli Stati dovranno inoltre garantire possibilità di ricorso alle vittime di discriminazione algoritmica e dovranno anche attivare strumenti di controllo indipendente per verificare la conformità alla convenzione. In oltre, dovranno sensibilizzare e informare il pubblico sulle tecnologie legate all'IA.

Per quello che riguarda gli aspetti che toccano la democrazia, gli Stati dovranno vigilare affinché l'IA non sia usata per danneggiare le istituzioni e i principi democratici di separazione dei poteri e di accesso alla giustizia. L'aspetto più critico, che è stato anche fonte di accese discussioni, ed è forse il limite di questa convenzione, riguarda la sua validità che tocca solo il settore pubblico e le sue aziende fornitrici di beni o servizi. Per quello privato, ogni Stato potrà decidere in modo autonomo, sce-

gliendo due opzioni: adottare la regolamentazione della convenzione in modo automatico anche per l'ambito privato o varare delle misure ad hoc per questo ambito per adeguarsi ai principi enunciati dalla convenzione.

Lei fa parte di un gruppo di persone esperte che ha l'obiettivo di elaborare delle raccomandazioni per gli Stati che sottoscriveranno la Convenzione, aperta alla firma il 5 settembre 2024 a Vilnius, in Lituania. Qual è lo scopo di queste raccomandazioni?

La Commissione per l'uguaglianza di genere (CEG) e il Comitato direttivo sull'antidiscriminazione, la diversità e l'inclusione (CDADI) del Consiglio d'Europa hanno incaricato un gruppo di persone esperte per elaborare le raccomandazioni da fornire al Comitato dei ministri, che è l'organo decisionale del Consiglio d'Europa ed è composto dai Ministri degli Affari esteri di tutti gli Stati membri. Queste raccomandazioni, ispirate allo studio sull'impatto dei sistemi di intelligenza artificiale in materia di parità e discriminazione² che ho redatto con Raphaële Xenidis per conto della CEG e del CDADI nel 2023, toccano differenti ambiti dove gli Stati possono agire per rendere operativa la Convenzione. C'è l'aspetto della prevenzione, della trasparenza e della responsabilità; quello importante dell'accesso alla giustizia; la questione della diversità delle comunità professionali, dell'inclusione e della promozione della parità e infine quello della partecipazione democratica e della sensibilizzazione dell'opinione pubblica. Nel processo di definizione di queste raccomandazioni si ritrova il modus



Ivana Bartoletti

«Si dovrebbe considerare ogni sistema basato sull'IA discriminatorio per definizione.»

«Siamo dunque di fronte a un problema sia tecnico, sia sociopolitico, la cui soluzione richiede una risposta che non sia solo tecnica.»

operandi del Consiglio d'Europa che prevede un dialogo con gli Stati membri, che a loro volta coinvolgono la società civile, i diversi gruppi d'interesse, oltre ai ricercatori e agli esperti sulla materia. Ovviamente il ruolo principale e decisivo spetta ai singoli Stati, ma si può dire che c'è un allineamento su quanto emerso nel nostro studio.

Quali sono i passi necessari per evitare che l'IA contrasti l'uguaglianza attraverso gli algoritmi?

È fondamentale una maggiore conoscenza sui bias algoritmici, cioè quegli errori sistematici di giudizio o di interpretazione, che portano a valutazioni sbagliate o poco oggettive. Questi bias emergono proprio dalla natura intrinseca dell'IA e dal suo funzionamento: l'IA usa dati che non sono neutri ma influenzati dal contesto e rispecchiano le disparità presenti nella società. Quando questi dati vengono immessi in una macchina per allenare un algoritmo a prendere delle decisioni o a formulare delle previsioni, si replica questa disparità e la si conferma con un software: è una softwarizzazione che codifica e perpetua le disuguaglianze esistenti nella realtà su una scala molto più grande. Un artefatto di intelligenza artificiale non è però solo frutto della tecnologia, perché i parametri con cui addestrare l'algoritmo sono scelti da persone. La domanda da porsi è semplice: chi decide di inserire determinati parametri in un sistema di IA? Chi compie questa scelta che ne esclude per forza altre? Questo processo decisionale, questa interazione fra la macchina e l'essere umano, può essere fonte di discriminazione ed è spesso difficile da individuare. Questa scelta è spesso operata da team di profes-

sionisti, che non hanno un background misto, non sono rappresentativi di tutti i gruppi sociali, tantomeno di quelli minoritari o svantaggiati. La mancanza di diversità e inclusione nei team di programmatori di sistemi IA è un fattore che gioca un ruolo importante. Siamo dunque di fronte a un problema sia tecnico, sia sociopolitico, la cui soluzione richiede una risposta che non sia solo tecnica, come spesso è accaduto in passato. Serve un approccio multidisciplinare, che inglobi le scienze sociali, il diritto, l'etica, le norme sulla non discriminazione, ma occorre anche la volontà politica, la volontà degli Stati, di volere da questi sistemi IA un risultato equo e non una fotocopia di una situazione diseguale.

Quali sono le difficoltà su questo cammino?

Quando si affronta questo tema, fa sempre capolino l'obiezione: ma se gli esseri umani sono portatori di un bias, che problema c'è se lo sono anche gli algoritmi? Il punto è che quando si inserisce la discriminazione in un software e la si automatizza, diventa poi complicato contrastarla. È anche difficile capire l'origine del bias, cioè identificare in quale tappa della pipeline – quella catena di elementi elaborati a cascata in un software – si produca. Inoltre, questi algoritmi agiscono all'interno di macchine, che vengono chiamate scatole nere, che imparano da sole e diventano incomprensibili anche ai professionisti del settore. Scatole nere verso le quali è complicato sollevare contestazioni da parte di chi si sente danneggiato, perché sono poco accessibili e protette dalla proprietà intellettuale.

Finora sono mancati gli strumenti legali e tecnici che possano aiutare a contrastare l'opacità di funzionamento e le discriminazioni dell'IA. A livello europeo, c'è ora la Legge sull'intelligenza artificiale (*Artificial Intelligence Act*) e il Regolamento generale per la protezione dei dati personali (in Svizzera la nuova Legge sulla protezione dei dati è entrata in vigore il 1° settembre del 2023), che pongono l'accento sulla *fairness*, cioè sull'equità e la correttezza delle decisioni prese da algoritmi. Questa *fairness* va costruita. Per esempio, se in una selezione del personale si vuole avere un esito equo dal punto di vista del genere, ma i dati a disposizione considerano poche donne, si possono introdurre dati sintetici, inventati, al fine di ottenere il risultato auspicato. Questo presuppone una chiara volontà di rendere il sistema di IA giusto e questa volontà non può che venire da una decisione da parte di chi produce questi sistemi. In questi anni, il Regolamento generale per la protezione dei dati personali è stato importante per appellarsi alla non trasparenza dell'algoritmo e dunque rifiutarsi di fornire il consenso per il trattamento dei dati personali. Al contrario, la normativa sulla discriminazione non è così efficace a contrastare il bias di correlazione prodotto dall'algoritmo, la cosiddetta *proxy discrimination*, che per esempio avviene quando si discrimina non solo perché un individuo è donna, ma anche perché naviga in determinati siti e perché si abita in un certo quartiere. Questa correlazione matematica che compie l'algoritmo, associando dati diversi, crea nuove categorie di discriminazione, che non sono quelle tradizionali (l'essere donna, l'essere disabile, l'appartenere a un gruppo religioso...), verso le quali la legge attuale non è così incisiva.

Dal punto di vista del quadro legale cosa servirebbe?

Uno degli elementi più forti che abbiamo proposto nel nostro studio è l'inversione dell'onere della prova. Si dovrebbe considerare ogni sistema basato sull'IA discriminatorio per definizione, a meno che i suoi sviluppatori non provino il contrario. Questo cambio di paradigma faciliterebbe il ricorso alla giustizia in caso di discriminazione e andrebbe a riequilibrare un'asimmetria di potere creata dai sistemi IA anche in termini di accesso alle informazioni e alla conoscenza sul loro funzionamento. Per

una persona normale è infatti un'impresa quasi impossibile raccogliere le prove di una discriminazione algoritmica. Questa presunzione di un bias discriminatorio si giustifica con l'alta frequenza di questo fenomeno e si dovrebbe applicare solo quando non ci sono delle misure preventive, come per esempio un'analisi di impatto, un audit che verifichi le procedure o una certificazione dell'algoritmo utilizzato. In questa architettura preventiva, dovrebbe essere presente un controllo pubblico da parte delle istituzioni di difesa dei diritti umani e degli enti contro la discriminazione, ma anche delle azioni positive per incentivare lo sviluppo di sistemi IA che soddisfino standard di qualità ed equità.

Il quadro è ora facilitato dall'adozione dall'*AI Act* dell'Unione europea, che prevede tutta una serie di requisiti e di obblighi e fra questi anche una valutazione di conformità sull'equità degli algoritmi.

Lei ha fondato e dirige Women Leading in AI³: di cosa si occupa questa associazione?

È nata nel 2018 ed è una rete di donne attive in tutti gli ambiti dell'intelligenza artificiale e quindi non solo in quello della programmazione. Raggruppamento professioniste provenienti dal settore tecnico, programmatrici, informatiche, scienziate, ma anche giornaliste, politiche, avvocate, esperte di etica e di diritti umani, donne attive nelle istituzioni pubbliche e nelle imprese private, ricercatrici e docenti universitarie. L'idea è quella di allargare il discorso sulla governance dell'IA e sul suo sviluppo responsabile, sulla definizione di linee guida, sugli strumenti per renderla un motore di innovazione e non di discriminazione. Abbiamo messo a punto un documento con dieci principi e altrettante raccomandazioni sulla regolamentazione dell'IA, che abbiamo indirizzato ai governi; vogliamo mobilitare la politica, ma anche sensibilizzare l'opinione pubblica, affinché sia meglio informata su questo importante tema. Crediamo nella costruzione di un'intelligenza artificiale che sostenga i nostri obiettivi e sia ancorata ai nostri valori umani.

«Occorre anche la volontà politica, la volontà degli Stati, di volere da questi sistemi IA un risultato equo e non una fotocopia di una situazione diseguale.»

Isabella Visetti è una giornalista RSI, conduce il magazine radiofonico *Mille voci* su Rete Uno e da tempo si occupa di temi legati alla parità di genere.

«Gli Stati dovranno inoltre garantire possibilità di ricorso alle vittime di discriminazione algoritmica.»

Note

- 1 Consiglio d'Europa: Council of Europe Framework Convention on Artificial Intelligence and Human Rights, Democracy and the Rule of Law (Convezione quadro del Consiglio d'Europa sull'intelligenza artificiale e i diritti umani, la democrazia e lo Stato di diritto), 2024. <https://rm.coe.int/1680afae3c> (consultato il: 25.06.2024).
- 2 Bartoletti, Ivana / Xenidis, Raphaële: Étude sur l'impact des systèmes d'intelligence artificielle, leur potentiel de promotion de l'égalité, y compris l'égalité de genre, et les risques qu'ils peuvent entraîner en matière de non-discrimination. Conseil de l'Europe, 2023. <https://rm.coe.int/study-on-the-impact-of-artificial-intelligence-systems-their-potential/1680ac99e3> (consultato il: 25.06.2024).
- 3 Women Leading in AI: <https://womenleadinginai.org/> (consultato il: 25.06.2024).

Abstracts

«Diskriminierungen durch Algorithmen sind nicht nur eine technische, sondern auch eine soziale und politische Frage»

Künstliche Intelligenz kann Gleichberechtigung und Inklusion untergraben. Eine Massnahme dagegen ist das *Rahmenabkommen des Europarats über die Künstliche Intelligenz, Menschenrechte, Demokratie und Rechtsstaatlichkeit*. **Ivana Bartoletti** ist Mit-herausgeberin der Studie des Europarats zu den Auswirkungen von KI-Systemen auf Gleichstellung und Diskriminierung, die Empfehlungen für fairere Algorithmen und KI-Systeme enthält. Die Probleme sind sowohl technisch als auch soziopolitisch und um sie zu lösen, braucht es einen multidisziplinären Ansatz. Auf gesetzlicher Ebene müssten alle KI-Systeme standardmässig als diskriminierend eingestuft werden, solange sie nicht das Gegenteil beweisen – dies wäre eine Umkehrung der Beweislast, die Opfern den Zugang zur Justiz erleichtern würde.

«La discrimination algorithmique n'est pas qu'une question technique, mais aussi sociale et politique»

L'intelligence artificielle risque de mettre fin aux espoirs d'égalité et d'inclusion. Pour y faire face, le Conseil de l'Europe a adopté la *Convention cadre sur l'intelligence artificielle, les droits de l'homme, la démocratie et l'État de droit*. **Ivana Bartoletti** est co-responsable de l'étude du Conseil de l'Europe sur l'impact des systèmes d'intelligence artificielle en matière d'égalité et de discrimination, qui formule des recommandations pour rendre les algorithmes et ces systèmes plus équitables. Il s'agit certes d'un problème technique, mais aussi sociopolitique, et il faut adopter une approche multidisciplinaire pour le résoudre. Sur le plan juridique, tout système d'IA devrait être considéré par défaut, sauf preuve du contraire, comme discriminatoire : un renversement de la charge de la preuve qui faciliterait l'accès à la justice pour les victimes.

« Nous devons assumer la responsabilité des préjugés qui sont au cœur des modèles algorithmiques »

Caitlin Kraft-Buchman promeut une approche humaniste et féministe dans le secteur de la technologie. Fondatrice du think tank Women at the Table et co-fondatrice de la coalition multidisciplinaire A+ Alliance, elle présente son engagement et sa vision d'une intelligence artificielle plus juste.

Portrait : Mathilda Golay

Un réseau de recherche féministe sur l'IA

En 2015, Caitlin Kraft-Buchman fonde Women at the Table, un think tank¹ destiné à promouvoir la représentation des femmes dans les sphères décisionnelles internationales. Cette expérience la pousse à s'intéresser aux algorithmes et, en 2019, elle co-fonde A+ Alliance, une coalition multidisciplinaire se concentrant sur la question de l'intelligence artificielle. À travers le financement de projets pilotes et d'appels à contributions, A+ Alliance se veut un « incubateur » de projets innovants et d'avancée technologique et sociale.

Concrètement, A+ Alliance vise à financer et développer des initiatives en lien avec l'IA. Régulièrement, des appels à projets sont lancés et un comité multidisciplinaire sélectionne les propositions qui se verront remettre ces fonds. Une attention toute particulière est donnée à la représentation des minorités, tant de la part des demandeur·se·s de fonds que dans les groupes qui bénéficieraient du projet fini. Cette partie se développe en plusieurs phases, allant de la proposition à l'implémentation d'un prototype de test sur le terrain. Par exemple, le prototype d'un système d'IA destiné à être utilisé dans les classes égyptiennes d'écoles communautaires – où les filles sont surreprésentées – est en cours de développement.² Cet outil permettrait aux enseignant·e·s d'évaluer la qualité du contenu généré par l'IA et le personnaliser en fonction de leurs objectifs pédagogiques et du programme d'études.³

Ainsi, les idées développées visent à améliorer la qualité de vie de groupes minoritaires, auxquels l'on ne pense pas d'emblée. Pour Caitlin Kraft-Buchman, l'intelligence artificielle doit être

féministe, et surtout, comme elle le dit elle-même, « pro-sociale », c'est-à-dire non seulement considérer les minorités souvent oubliées, mais être pensée en termes de contributions positives aux vies de tous et toutes : « Ces notions d'inclusion et de transformation impliquent que les communautés concernées doivent vraiment être au cœur des processus de développement, pour que nous créions une technologie dont nous avons effectivement besoin. »

Une approche multidisciplinaire

Afin de créer des outils et programmes qui permettront de développer une IA inclusive, il faut amener une variété de points de vue. La co-fondatrice insiste sur le fait que ce qui fait la différence, c'est avant tout « une approche multidisciplinaire, avec des personnes progressistes qui se soucient profondément des autres ». Au-delà de la représentation de différentes disciplines, il est également central que les personnes qui ont fait l'expérience des conséquences de l'utilisation des algorithmes soient représentées dans la discussion et « participent à la cocréation de solutions ».

Pour Caitlin Kraft-Buchman, lorsqu'il s'agit de créer des outils ayant un impact positif et qui permettent une avancée sociale réelle et juste, l'expérience vécue vaut autant que l'expertise technique. Ainsi, les personnes concernées doivent faire partie de la conversation, et être incluses dans les projets de développement. Et cela n'est possible que si plusieurs groupes sont entendus et participent à la conception de ces nouveaux outils qui ont une réelle utilité pour eux : « Il ne s'agit pas uniquement de créer des applications pour pouvoir se faire livrer des pizzas à domicile. » Cette exigence de la



Caitlin Kraft-Buchman

« Nous avons maintenant la possibilité d'un nouveau choix, et une opportunité de mettre en place de nouveaux systèmes. »

« À quoi ressemblerait une IA qui non seulement ne discrimine pas, mais qui permet de renforcer l'autonomie des communautés ? »

multidisciplinarité n'est pas sans rappeler le modèle suisse, dont l'entrepreneuse américaine s'est inspirée. Dans les deux cas, la position de différents groupes est entendue, et le résultat, que ce soit une politique publique ou un produit technologique, en bénéficie.

Le potentiel de l'IA

Les modèles financés par A+ Alliance ne sont pas technologiquement révolutionnaires ; leur intérêt repose principalement dans la manière dont les outils déjà à disposition sont utilisés, et dans le lien qu'ils permettent de créer entre qui les construit et qui peut en bénéficier. Pour Caitlin Kraft-Buchman, nous n'exploitons actuellement pas le plein potentiel de ces outils, actuellement développés par un groupe de personnes d'une démographie très monolithique, qui les fabrique à leur propre image. Or, selon elle, la promesse de l'IA est de permettre à différentes personnes de se retrouver et de créer quelque chose de nouveau. Mais cette promesse n'est réalisable qu'à une condition : que nous la gardions en tête lorsque nous créons de nouveaux outils avec l'IA. Autrement dit, que cette création se fasse de manière intentionnelle, avec les bonnes questions posées en amont. Car, nous avertit Caitlin Kraft-Buchman, « à moins d'en avoir l'intention et de créer les processus, les nouvelles idées et les nouvelles façons de penser autour de cette promesse, nous allons juste finir par reproduire le statu quo. Nous avons maintenant la possibilité d'un nouveau choix, et une opportunité de mettre en place de nouveaux systèmes. »

Changer le système pour plus d'impact

Un autre aspect du travail de Caitlin Kraft-Buchman est celui à niveau international. L'organisation Women at the Table s'engage pour que les considérations de genre et d'égalité soient prises en

compte dans les documents légiférant l'utilisation de l'IA. La fondatrice a notamment travaillé à l'élaboration d'un plan d'action sur le sujet de la technologie et de l'innovation qui fut l'armature de la 67^e session annuelle de la Commission de la condition de la femme des Nations Unies en 2023, session qui fut la première à traiter des thèmes de la technologie et de l'innovation.^{4,5}

Pour elle, l'intérêt de ce travail international réside dans l'impact qu'il permet d'avoir. Elle considère les légiférations internationales comme ayant un pouvoir transformatif très élevé, et, de ce fait, comme pouvant réellement changer la manière dont nous concevons et utilisons l'IA. Sa stratégie consiste ainsi à agir au niveau le plus haut possible et ainsi « changer le système ». Une approche pour elle résolument féministe : « Nous apprenons que les femmes doivent se modeler elles-mêmes pour s'adapter au système, ce qui peut être très utile pour elles. Et toutes nous avons dû le faire et le faisons probablement chaque jour. Cependant, maintenant que nous sommes là, ce que nous devons faire, c'est changer le système pour que les femmes, dans toute leur diversité, et les autres personnes puissent être elles-mêmes, et que le système puisse répondre aux besoins de tout le monde. »

Opérer au niveau international lui permet également d'aborder la question de la responsabilité. Les algorithmes ayant été élaborés par des êtres humains dans un contexte donné, ils contiennent les biais de société déjà présents dans les données qui ont été utilisées pour les nourrir, provenant notamment d'internet. « Mais pas tout le monde n'est présent sur internet », rappelle Caitlin Kraft-Buchman. Les algorithmes sont créés à l'image d'un certain groupe et opèrent en conséquence d'une certaine

manière. Ils ne sont ainsi pas représentatifs de la majorité des personnes. « Mais au lieu de simplement lancer « oh c'est l'algorithme qui l'a dit », je pense que nous devons assumer la responsabilité des préjugés qui sont au cœur des modèles algorithmiques. » Et cette responsabilité incombe à toute personne pouvant influencer la manière dont les outils algorithmiques sont développés.

Il faut se concentrer sur les solutions

Assumer la responsabilité signifie également chercher de nouvelles solutions. Pour la co-fondatrice d'A+ Alliance, pas question de continuer à simplement reproduire ce qui existe déjà. Au travers de sa participation à la mise en place de standards et de la création d'outils basés sur l'IA, elle prône par l'exemple une attitude proactive et créative. Porter une attention toute particulière aux modèles en cours de création est également une réponse à la question de la mitigation des stéréotypes dans les algorithmes. Une fois que les modèles algorithmiques contiennent des stéréotypes, il est très difficile d'éliminer ces derniers. La solution, alors, est plutôt de considérer les outils qui n'ont pas encore été créés, et s'assurer qu'ils seront meilleurs que leurs prédécesseurs. Caitlin Kraft-Buchman propose par exemple de se concentrer sur des modèles plus petits, c'est-à-dire qui ne peuvent pas être utilisés à large échelle, mais qui peuvent être plus soignés et précis, permettant d'éviter le piège des stéréotypes et de mieux servir les groupes auxquels ils sont destinés.

Se mobiliser pour la recherche de solutions n'est pas seulement une question de progrès, mais également d'égalité, avertit la CEO : « Les femmes et tous les autres groupes marginalisés ont été exclus de la création d'outils, de l'élaboration de règles, de la prise de décision. Si nous ne commençons pas à créer des outils qui fonctionnent pour nous, nous serons toujours confiné-e-s aux structures existantes. Nous devons donc maintenant vraiment réfléchir à la manière dont nous pouvons prendre part à ces processus. » Remettre les choses dans les mains des personnes directement concernées dans l'élaboration de l'IA et dans sa mise en œuvre permet de leur donner une voix, de leur donner

l'opportunité de développer leurs propres idées, et ainsi de rendre les processus plus démocratiques. Si l'IA est utilisée dans de nombreux domaines avec tous les risques que cela comporte, les principales victimes potentielles doivent avoir leur mot à dire. De mettre ces dernières au cœur de projets de développement d'IA permet de démocratiser le processus : « En tant que féministes, militant-e-s, simples citoyen-ne-s, nous ne serons jamais admises dans ce monde » rappelle Caitlin Kraft-Buchman. « D'une certaine manière, nous renonçons donc à notre pouvoir d'action. C'est pourquoi je pense que ces petites expériences simples sont si importantes. En fin de compte, je pense qu'il s'agit effectivement d'une question de démocratie et de la création d'outils qui permettent à cette dernière de fonctionner. »

Que ce soit dans le travail de Women at the Table ou de A+ Alliance, dans son engagement dans la promotion de projets innovants et de travail à l'international pour faire légiférer l'utilisation de l'IA afin que les minorités et les processus démocratiques soient pris en compte, Caitlin Kraft-Buchman a toujours la question du futur devant les yeux. Que souhaitons-nous pour le futur ? Quelle sorte de futur ? À quoi ressemblerait une IA qui non seulement ne discrimine pas, mais qui permet de renforcer l'autonomie des communautés ? En matière de solutions, elle considère qu'il faut désormais se poser ces questions avant qu'il ne soit trop tard. Est-ce possible de louper le coche ? Oui, nous dit-elle : « C'est un moment dangereux ; si nous continuons de faire comme nous avons toujours fait et nous continuons de produire la même chose à une cadence plus rapide, alors je pense que c'est dangereux pour la démocratie. Cela va creuser les écarts et inégalités qui existent à tous les niveaux, et cela va nuire à la cohésion sociale. » Et d'ajouter : « Et il n'y a pas de démocratie sans les femmes. »

Mathilda Golay est stagiaire à la CFQF. Elle est diplômée en philosophie et en sociologie de l'Université de Fribourg et de l'Université d'Amsterdam.

« Si nous continuons de faire comme nous avons toujours fait et nous continuons de produire la même chose à une cadence plus rapide, alors je pense que c'est dangereux pour la démocratie. »

Notes

- 1 Un think tank est un regroupement d'expert-e-s dont le but est de développer de la recherche et des réflexions autour d'un certain sujet. Wikipedia : Think tank, 2024. https://fr.wikipedia.org/wiki/Think_tank (consulté le : 21.05.2024).
- 2 Soudi, Marwa et al. : Generative AI-Based Tutoring System for Upper Egypt Community Schools. In : Proceedings of the 2023 Conference on Human Centered Artificial Intelligence : Education and Practice (HCAIep 23), 2023. <https://doi.org/10.1145/3633083.3633085>.
- 3 A+ Alliance : IdeasGym is prototyping Explainable AI-Based Tutoring in Upper Egypt rural schools, 2023. <https://aplusalliance.org/ideasgym-is-prototyping-explainable-ai-based-tutoring-in-upper-egypt-rural-schools/> (consulté le : 18.05.2023).
- 4 UNWomen : 67CSW, 2023. <https://www.unwomen.org/en/how-we-work/commission-on-the-status-of-women> (consulté le : 20.05.2024).
- 5 La Commission de la Condition de la Femme, ou CSW, est un organe intergouvernemental des Nations Unies dédié à promouvoir l'égalité de genre et l'autonomisation des femmes.

« Les personnes concernées doivent faire partie de la conversation, et être incluses dans les projets de développement. »

Abstracts

«Wir müssen Verantwortung übernehmen für die Vorurteile, die in algorithmischen Modellen vorhanden sind»

Caitlin Kraft-Buchman ist Mitgründerin des Think-tank Women at the Table und der multidisziplinären Allianz A+ Alliance. Die beiden Organisationen wollen auf internationaler Ebene den feministischen Einfluss in Schlüsselbereichen wie der Technologie verstärken. Um dieses Ziel zu erreichen, setzen sie auf eine starke Präsenz an multilateralen Treffen und bei der Entwicklung von KI-Projekten. Caitlin Kraft-Buchman erläutert in diesem Porträt ihr humanistisches Engagement und spricht sich für einen proaktiven, multidisziplinären und kreativen Ansatz im Umgang mit den durch Künstliche Intelligenz entstandenen Herausforderungen aus.

«Dobbiamo assumerci la responsabilità dei pregiudizi alla base dei modelli algoritmici»

Caitlin Kraft-Buchman ha fondato il think tank Women at the Table e co-fondato la coalizione multidisciplinare A+ Alliance, due organizzazioni che mirano a promuovere un'influenza femminista a livello internazionale in settori chiave come quello tecnologico. A tale scopo, esse puntano su una forte presenza agli incontri multilaterali e sullo sviluppo di progetti legati all'intelligenza artificiale. In questo ritratto, presenta il suo impegno umanista e raccomanda un approccio proattivo, multidisciplinare e creativo alle sfide poste dall'intelligenza artificiale.

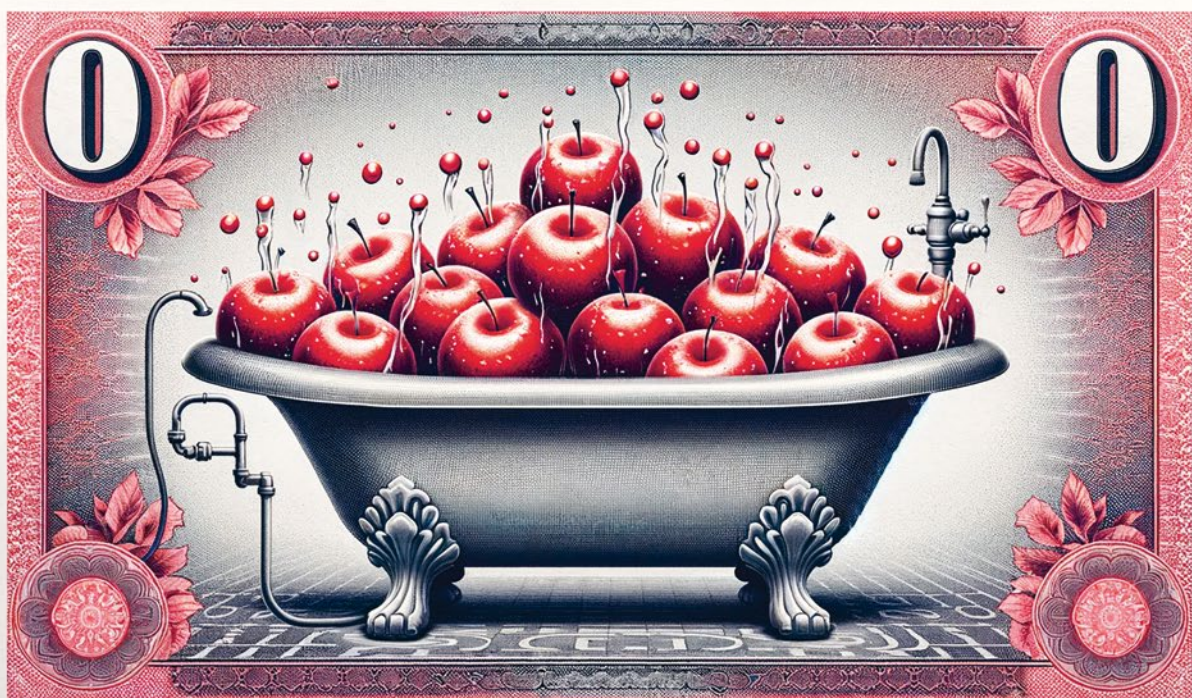
Swiss Money

DALL-E, can you make a Swiss banknote?

I cannot create or generate images of real currency, such as Swiss banknotes, for legal reasons and to respect copyright. However, I can help you design a fictional banknote inspired by Swiss style or cultural elements if you wish. Let me know what you would like to see on this fictional banknote!

Yes, ok, let's create a banknote with a bathtub full of apples, and the number 0.

Here is a fictional Swiss banknote design in a rectangular format, featuring a whimsical motif of a clawfoot bathtub filled with shiny red apples, now presented with a monochrome background for a starker and more modern look. The bathtub and apples remain vibrant and colorful, creating a striking contrast against the simple, single-color backdrop. The design retains its artistic elements with intricate patterns bordering the banknote and the denomination displayed as a large number "0" in an elegant, large font. This version emphasizes simplicity and focus, blending playful creativity with a minimalist aesthetic.



«Öffentliche Verzeichnisse sind ideal, um über die Risiken und Chancen von KI zu diskutieren»

Für **Liliane Obrecht**¹ ist klar: Damit beim Einsatz von Künstlicher Intelligenz keine Diskriminierungen durch den Staat entstehen, braucht es mehr Transparenz über die Anwendung von KI-Systemen in öffentlichen Verwaltungen. Die Juristin plädiert im Gespräch für die Einführung von allgemein zugänglichen Verzeichnissen und sagt, wer schon heute mit gutem Beispiel vorangeht.

Interview: Isabel Knobel

Liliane Obrecht, Sie forschen zu Rechtsfragen beim Einsatz von Künstlicher Intelligenz in öffentlichen Verwaltungen in der Schweiz. Was ist KI eigentlich?

Die Bezeichnung KI wird häufig als eine Art Sammelbegriff für verschiedene Technologien und Methoden verwendet, die auf selbständig lernenden Algorithmen beruhen. Dazu zählen z.B. maschinelle Lernverfahren, aber auch die natürliche Sprachverarbeitung, die seit dem Aufkommen von ChatGPT besonders populär ist. In der Schweiz gibt es bis heute keine rechtlich verbindliche Definition. Einige Forschende plädieren für die Nutzung eines alternativen Begriffs, nämlich den der algorithmischen Systeme. Viele Risiken, die mit selbstlernenden Systemen einhergehen, sind auch beim Einsatz von Algorithmen, die nicht zwingend von selbst lernen, vorhanden. Mit diesem Begriff wären sie auch eingeschlossen. Zudem ist der Begriff technologie-neutraler und im öffentlichen Diskurs weniger negativ konnotiert als KI.

Wie werden KI und algorithmische Systeme in öffentlichen Verwaltungen eingesetzt?

Es gibt viele verschiedene Einsatzbereiche, doch wir wissen heute nicht genau, wo welche Systeme eingesetzt werden. Gut sichtbar ist der Einsatz von Chatbots². Das sind Onlinedialogsysteme, die natürliche Sprache verarbeiten und dadurch mit Menschen kommunizieren können. Chatbots können – müssen aber nicht – KI nutzen, um Fragen zu verarbeiten und automatisierte Antworten zu erstellen. Die Eidgenössische Stiftungsaufsicht ESA hat beispielsweise im Januar 2024 den Chabot *Esi* lan-

ciert, auf Basis von ChatGPT 3.5 turbo. Er beantwortet Fragen rund um das schweizerische Stiftungswesen. Auch die Kantone setzen vermehrt auf Bots. Wer z.B. beim Strassenverkehrsamt des Kantons Aargau ausserhalb der Öffnungszeiten anruft, wird mit dem Voicebot *Vicky* verbunden. *Vicky* versteht auch Mundart, lernt laufend dazu und soll künftig die Hälfte aller Anrufe beantworten können.

Warum setzen Behörden vermehrt auf maschinelle Anwendungen?

Ich denke, sie erhoffen sich davon in erster Linie eine Effizienzsteigerung. Der demografische Wandel wird dazu führen, dass viele Stellen unbesetzt bleiben und diese Lücken müssen irgendwie geschlossen werden. Algorithmische Systeme und KI bieten sich an, standardisierte, repetitive Aufgaben zu übernehmen. Es geht dabei nicht darum, Mitarbeitende zu ersetzen, sondern sie zu entlasten, damit sie mehr Zeit haben, sich mit komplexeren Fragen zu beschäftigen. Die Gefahr, dass KI den Menschen ihren Arbeitsplatz wegnimmt, wird aus meiner Sicht überzeichnet.

Der Privatsektor ist häufig schneller und innovativer als der öffentliche Sektor. Ist das bei KI auch so?

Der Staat untersteht strengeren Regeln beim Einsatz von KI. Er muss sich an das Legalitätsprinzip halten und die Grundrechte respektieren. Das heisst, er darf nur handeln, wenn eine ausreichende gesetzliche Grundlage existiert. Überdies besteht eine sehr sensible Beziehung zwischen Staat und Bevölkerung. Der Staat handelt gegenüber den Individuen hoheitlich, das heisst, er kann



Liliane Obrecht

«Es besteht ein Machtungleichverhältnis, das potenziell in Machtmissbrauch kippen kann.»

«Mitarbeitende von Behörden sollten so geschult werden, dass sie diskriminierende maschinelle Entscheide erkennen können.»

einseitig über Rechte und Pflichten der Betroffenen bestimmen. Wenn wir z. B. unsere Steuererklärung ausfüllen, können wir nicht aussuchen, zu welcher Behörde wir gehen. Wenn diese Behörde KI einsetzt, sind wir dieser Tatsache «ausgeliefert». Es besteht ein Machtungleichverhältnis, das potenziell in Machtmissbrauch kippen kann. Deshalb ist es gut, dass der Staat höheren Anforderungen untersteht und es entsprechend auch länger dauert, bis öffentliche Verwaltungen KI einsetzen können.

Sie sprechen die Risiken von KI an. Welche sind das?

Ein zentrales Risiko ist die Benachteiligung von bestimmten Bevölkerungsgruppen. Dazu gehören etwa Frauen, aber auch Menschen mit Behinderung oder People of Color. KI-Systeme werden mit sehr grossen Datensätzen trainiert. Es kann sein, dass diese Datensätze veraltete Daten enthalten oder nicht vollständig sind. Werden zum Beispiel in einem Bewerbungsverfahren für eine hochqualifizierte Stelle Daten der letzten Jahrzehnte eingespeist, schlägt das System möglicherweise vor, einen Bewerber anstelle einer Bewerberin zu wählen. Es erachtet Männer als besser qualifiziert, weil diese bis dato häufiger höhere Positionen besetzt haben. Daten aus der Vergangenheit können also dazu führen, dass KI-Systeme in der Gegenwart Diskriminierungen perpetuieren, gegen die die Gesellschaft gerade versucht vorzugehen. In Österreich z. B. setzte der Arbeitsmarktservice (AMS) einen auf maschinellen Lernverfahren beruhenden Algorithmus ein, der Arbeitssuchende nach ihrer Vermittelbarkeit einstuft und so indirekt über ihre Jobchancen entschied. Gegen dieses Projekt wird seit Jahren gerichtlich vorgegangen.³

Gibt es gesetzliche Bestimmungen, die vor solchen Diskriminierungen schützen?

Die Bundesverfassung statuiert in Artikel 8 Absatz 2 das Diskriminierungsverbot und listet eine Reihe von Merkmalen wie Geschlecht, Herkunft, Rasse, Alter, Sprache oder soziale Stellung auf. Im Kontext von KI können insbesondere sogenannte indirekte Diskriminierungen vorkommen, wenn neutral erscheinende Kriterien zu faktischen Nachteilen für Angehörige einer dieser geschützten Gruppen führen. Es ist wichtig und gut, dass wir diesen Rechtsrahmen bereits haben, aber es wird zusätzliche Rechtsgrundlagen brauchen, damit das Diskriminierungsverbot bei KI-Anwendungen zum Tragen kommt. Eine Diskriminierung lässt sich häufig nicht an einem Einzelentscheid feststellen, sondern er-

gibt sich erst im grösseren Kontext aus dem Einsatz dieser Systeme. Deswegen müssen sie ausreichend reguliert und kontrolliert werden.

Wie gross ist die Gefahr von Diskriminierung durch den KI-Einsatz von Schweizer Behörden tatsächlich?

Meines Wissens werden in der Bundesverwaltung bislang kaum heikle algorithmische Systeme eingesetzt, auch weil die notwendigen Rechtsgrundlagen fehlen. Das kann aber nicht abschliessend beurteilt werden, weil es eben noch keine Übersicht gibt, wo KI überall eingesetzt wird. Klar ist aber, dass, wenn z. B. Steuerbehörden in Verwaltungsverfahren KI einsetzen möchten, es dafür neue Rechtsgrundlagen braucht. In unserem demokratischen Rechtsstaat dauert es relativ lange, bis solche geschaffen sind. Das ist in diesem Kontext gut. Die digitale Transformation des Staates beruht auf einem Vertrauensverhältnis zwischen Bevölkerung und Staat. Schnelle Innovation bringt dem Staat nichts, wenn die Bevölkerung ihm nicht vertraut.

Welche gesetzlichen Anpassungen schlagen Sie für die Schweiz vor, um Diskriminierungen bei einem KI-Einsatz zu verhindern?

Transparenz kann ein wirksames Mittel sein, um Diskriminierungen zu adressieren. Die Einführung von verpflichtenden öffentlichen Verzeichnissen über den Einsatz von algorithmischen Systemen – und eben nicht nur KI – in der öffentlichen Verwaltung ist ein wichtiger erster Schritt. Der Bund, konkret der Bundesrat inklusive öffentlicher Verwaltung, hat eine aktive Informationspflicht, die Bevölkerung über seine Tätigkeiten aufzuklären (Artikel 180 Absatz 2 Bundesverfassung). Meiner Meinung nach schliesst das auch Informationen über die Verwendung von KI-Systemen ein, also z. B. die Datengrundlage, Risiken für die Grundrechte und allfällige durchgeführte Kontrollen. Sowohl die breite Öffentlichkeit als auch Expertinnen und Experten müssen Zugang zu diesen maschinenlesbaren Verzeichnissen haben. Die Informationen können dabei unterschiedliche Spezifizierungsgrade enthalten. Für die Wissenschaft sind z. B. die technischen Details sehr interessant. Wichtig ist auch, dass nicht jedes einzelne System erfasst werden muss. Es wäre kaum zielführend, wenn jede administrative Hilfstätigkeit, die auf KI basiert, in einem solchen Verzeichnis aufgeführt würde. Aber es sollten jene Systeme offengelegt werden, die Auswirkungen auf die Menschen haben.

Gibt es Verwaltungen in der Schweiz, die ihre KI-Nutzung offenlegen?

Auf Bundesebene führt das Kompetenznetzwerk für künstliche Intelligenz CNAI eine Projektdatenbank mit einigen in der Bundesverwaltung umgesetzten KI-Projekten. Damit soll die Öffentlichkeit über die Tätigkeiten der Bundesverwaltung im Bereich KI informiert werden. Aktuell sind dort rund 70 Projekte aufgelistet, allerdings erst auf freiwilliger Basis. Der Bundesrat prüft, ob er eine Meldepflicht einführen will.⁴ Diesbezüglicher Vorreiter ist der Kanton Zürich: Als erster Kanton möchte er eine Rechtsgrundlage für ein verpflichtendes öffentliches Verzeichnis über den Einsatz von KI-Systemen schaffen, die die Grundrechte der Bevölkerung betreffen. Der Entwurf des Artikels ist im revidierten kantonalen Informations- und Datenschutzgesetz verankert⁵ und geht massgeblich auf das Engagement der zivilgesellschaftlichen Organisation AlgorithmWatch zurück. Auch im Kanton Basel-Stadt gab es bereits einen entsprechenden parlamentarischen Vorstoss.⁶

Beeinflussen auch internationale Entwicklungen die Rechtslage in der Schweiz?

Im März 2024 hat das Europäische Parlament den *AI Act* verabschiedet, den ersten umfassenden Rechtsrahmen zur Regulierung von KI.⁷ Als Nicht-EU-Mitglied ist die Schweiz zwar nicht zur Übernahme verpflichtet, aber für den privaten Sektor ist er in Bezug auf die Markttanschlussfähigkeit sehr relevant. Zudem hat der Europarat, in dem die Schweiz Mitglied ist, im Mai 2024 eine KI-Konvention verabschiedet.⁸ Es wird also noch einiges auf uns zukommen. Wir stehen bei der Regulierung von KI erst am Anfang.

Was braucht es neben Transparenz und neuen Rechtsgrundlagen?

KI und algorithmische Systeme entwickeln sich laufend weiter. Deshalb ist es wichtig, sie nicht nur transparent zu machen und bei der Inverkehrsetzung zu überprüfen, sondern sie laufend zu analysieren. Auch den menschlichen Umgang mit KI müssen wir verbessern, hier fehlt es noch an Know-how. Mitarbeitende von Behörden sollten so geschult werden, dass sie diskriminierende maschinelle Entscheide erkennen können. Besonders wichtig ist das bei teilautomatisierten Systemen, die keine Entscheidungen fällen, sondern den Mitarbeitenden Vorschläge unterbreiten. Handlungsempfehlungen sollten stets geprüft und nicht blind

übernommen werden. Darüber hinaus brauchen wir eine breite gesellschaftliche Auseinandersetzung mit dem Thema. Ich glaube, dass das Gespräch über öffentliche Verzeichnisse der perfekte Anlass ist, sowohl über die Risiken als auch die Chancen von KI zu diskutieren. Hier nehmen zivilgesellschaftliche Organisationen eine wichtige Rolle ein. Auch die Wissenschaft kann mit interdisziplinärer Forschung und Expertise in unterschiedlichen Bereichen viel beitragen.

Kann KI von Diskriminierung betroffene Personen auch unterstützen und schützen?

Hier sprechen Sie ein sehr wichtiges Thema an. Der Einsatz von KI kann auch dazu führen, dass Diskriminierungen erst erkennbar gemacht und behoben werden. Im Beispiel der Bewerbungsverfahren könnte KI historisch verzerrte Datensätze aufdecken. Und wenn Datengrundlagen in öffentlichen Verzeichnissen dargelegt und erläutert werden müssen, ist das eine Chance für besonders vulnerable Gruppen, z.B. Frauen oder Menschen mit Behinderungen, vermehrt berücksichtigt und repräsentativ abgebildet zu werden. Es besteht aber viel Forschungsbedarf in diesem Bereich, da er bisher nicht prioritär behandelt wurde. Mit dem breit angelegten Forschungsprojekt *Nachvollziehbare Algorithmen: Ein Rechtsrahmen für den Einsatz von Künstlicher Intelligenz*,⁹ das von der Stiftung Mercator unterstützt wird, versuchen wir einen Beitrag zu leisten.

Abschliessend noch der Blick in die Zukunft: Wie gehen wir in einer idealen Welt mit KI um?

Das ist eine schwierige Frage. Es ist wichtig, dass wir ein Gleichgewicht finden zwischen Innovation und Schutz von Persönlichkeit und Grundrechten, da sind sich mittlerweile alle einig. Eine Regulierung ist notwendig, das ist aus rechtlicher Sicht klar. Wie genau diese umgesetzt werden wird, wissen wir heute noch nicht. Der Bundesrat hat in Aussicht gestellt, bis Ende 2024 eine Auslegeordnung zu einer allfälligen KI-Regulierung zu präsentieren.¹⁰

«Wir wissen heute nicht genau, wo welche Systeme eingesetzt werden.»

Isabel Knobel ist Politologin und sucht als Projektleiterin beim Migros-Pionierfonds mutige Ideen für eine zukunfts-fähige Gesellschaft. Sie schreibt regelmässig für die Fachzeitschrift «Frauenfragen».

«Die Gefahr, dass KI den Menschen ihren Arbeitsplatz wegnimmt, wird überzeichnet.»

Anmerkungen

- 1 Liliane Obrecht ist Doktorandin an der Professur von Prof. Nadja Braun-Binder, wissenschaftliche Mitarbeiterin bei der fakultären Forschungsstelle für Digitalisierung in Staat und Verwaltung e-PIAF der Juristischen Fakultät der Universität Basel und Vorstandsmitglied bei AlgorithmWatch Schweiz.
- 2 Vom Englischen to chat = plaudern; robots = Roboter.
- 3 Pflügl, Jakob: Entscheidet die KI über Jobs? Höchstgericht lässt AMS-Algorithmus erneut prüfen. Der Standard, 10.02.2024. <https://www.derstandard.at/story/3000000206845/entscheidet-die-ki-ueber-jobs-hoehstgericht-laesst-ams-algorithmus-erneut-pruefen> (abgerufen am: 24.06.2024).
- 4 Po. Schlatter 23.3566: «Vollständigkeit des Transparenzregisters für den Einsatz von künstlicher Intelligenz durch die Bundesverwaltung sicherstellen», 04.05.2023. <https://www.parlament.ch/de/ratsbetrieb/suche-curia-vista/geschaeft?AffairId=20233566> (abgerufen am: 24.06.2024).
- 5 Kanton Zürich: Kanton Zürich modernisiert Gesetz über Information und Datenschutz. [Medienmitteilung], 04.08.2023. <https://www.zh.ch/de/news-uebersicht/medienmitteilungen/2023/08/kanton-zuerich-modernisiert-gesetz-ueber-information-und-datenschutz.html> (abgerufen am: 24.06.2024).
- 6 Grosser Rat des Kantons Basel-Stadt: Geschäft 23.5474, «Anzug der Justiz-, Sicherheits- und Sportkommission betreffend Transparenz über den Einsatz von algorithmischen Systemen in der Verwaltung», 18.10.2023. <https://grosserrat.bs.ch/dokumente/100405/000000405630.pdf> (abgerufen am: 24.06.2024).
- 7 <https://artificialintelligenceact.eu/de/>
- 8 Europarat: Europarat verabschiedet ersten internationalen Vertrag über künstliche Intelligenz, 17.05.2024. <https://www.coe.int/de/web/portal/-/council-of-europe-adopts-first-international-treaty-on-artificial-intelligence> (abgerufen am: 24.06.2024).
- 9 Universität Basel: Nachvollziehbare Algorithmen: ein Rechtsrahmen für den Einsatz von Künstlicher Intelligenz. e-PIAF. <https://ius.unibas.ch/de/e-piaf/nachvollziehbare-algorithmen/> (abgerufen am: 24.06.2024).
- 10 Bundesrat: «Bundesrat prüft Regulierungsansätze für Künstliche Intelligenz». [Medienmitteilung], 22.11.2023. <https://www.admin.ch/gov/de/start/dokumentation/medienmitteilungen.msg-id-98791.html>.

Abstracts

« Les registres publics des algorithmes utilisés sont idéaux pour discuter des risques et des opportunités de l'IA »

Chercheuse à l'Université de Bâle, la juriste **Liliane Obrecht** étudie les questions juridiques posées par l'utilisation de l'intelligence artificielle dans les administrations publiques. Dans cette interview, elle explique pourquoi l'établissement de registres publics des algorithmes utilisés par une administration est une première étape importante vers la conception d'un cadre juridique cohérent pour l'intelligence artificielle. Améliorer la transparence peut aider à détecter et contrer les discriminations générées par des jeux de données présentant des distorsions d'origine historique. Pour elle, l'IA ne présente pas que des risques, mais elle offre aussi des opportunités au secteur public. La chercheuse prône un large débat de société sur cette question.

«I registri pubblici sono ideali per discutere dei rischi e delle opportunità dell'IA»

La giurista **Liliane Obrecht** è ricercatrice presso l'Università di Basilea sulle questioni giuridiche legate all'uso dell'intelligenza artificiale nell'amministrazione pubblica. Intervistata, spiega perché i registri pubblici dei sistemi algoritmici utilizzati da un'amministrazione siano un primo passo importante verso un quadro legislativo coerente per l'intelligenza artificiale. Una maggiore trasparenza può contribuire a riconoscere ed eliminare le discriminazioni dovute per esempio a set di dati storicamente distorti. Oltre ai rischi, Liliane Obrecht vede anche opportunità per l'intelligenza artificiale nel settore pubblico e chiede un dibattito sociale più ampio su questo tema.

Wo steht die Schweizer Politik im Bereich KI? Das Co-Präsidium von Parldigi liefert Antworten

SP-Nationalrätin **Min Li Marti** und SVP-Nationalrat **Franz Grüter** bilden das Co-Präsidium der parlamentarischen Gruppe Parldigi, die sich mit dem Thema digitale Nachhaltigkeit befasst. Im Interview schätzen sie die Chancen und Risiken von KI-Tools ein und diskutieren den Regulierungsbedarf in der Schweiz. Sie zeigen auf, wie sie parteiübergreifend Mehrheiten finden, die Meinungen beim Thema KI und Diskriminierung aber doch auseinandergehen.

Interview: Nadja Senn

Frau Marti, Herr Grüter, Sie sind beide seit 2015 Mitglied bei Parldigi. Was hat Sie dazu bewegt, sich in der parlamentarischen Gruppe zu engagieren?

Min Li Marti: Da mich die Auswirkung der Technologie auf unsere Gesellschaft und unser Leben immer schon interessiert hat, bin ich Parldigi¹ gleich nach meiner Wahl beigetreten. Weil Digitalisierung ein politisches Nischenthema ist, sind die Debatten weniger ideologisch festgefahren und es ist möglich, überparteiliche Lösungen zu finden. Zudem fasziniert mich, dass sich oft neue Themengebiete und Herausforderungen ergeben, zu welchen noch keine Lösungsansätze vorhanden sind.

Franz Grüter: Ich kann mich hier anschliessen. Die technologische Entwicklung spielt zwar im Leben aller Menschen eine wichtige Rolle, führt in der Politik aber weiterhin ein Schattendasein. Bei Parldigi ist jeweils mindestens ein Vertreter oder eine Vertreterin pro Partei dabei. So kann die Gruppe digitale Themen aus verschiedenen Blickwinkeln beleuchten und stösst damit auf breite Resonanz in der Verwaltung, der Wirtschaft und in Organisationen, die sich mit der Digitalisierung befassen.

Können Sie Beispiele nennen für diese gut funktionierende parteiübergreifende Allianz?

Franz Grüter: Das beste Beispiel hierfür ist die Elektronische Identität (E-ID). Obwohl diese 2021 in der Volksabstimmung scheiterte, gelang es mittels einer parteiübergreifenden Allianz, das Thema erneut aufzunehmen. So konnte eine Vorlage durchs Parlament gebracht werden, die nicht mehr nur privatwirtschaftlich orientiert war. Dass nun an

deren Umsetzung gearbeitet wird, ist dem Umstand zu verdanken, dass nicht ideologisch, sondern sach- und themenorientiert gearbeitet wurde. Dieses Vorgehen ist auch dank unserer gut funktionierenden Zusammenarbeit im Co-Präsidium möglich. Ich möchte aber festhalten, dass wir unsere Erfolge nicht an neuen Gesetzen oder der Anzahl Vorstösse messen. Unser Ziel ist es, neue Themen aufzugreifen und deren gesellschaftliche Auswirkungen zu thematisieren. Wir beleuchten Themen oft kontrovers, laden Fachpersonen ein und diskutieren, um das Parlament bei der Meinungsbildung zu unterstützen.

Min Li Marti: Neben dem elektronischen Patientendossier werden uns in Zukunft digitale Plattformen, Algorithmen und Hatespeech beschäftigen. Unter letzterem Problem leiden wir Politikerinnen und Politiker fraktionsübergreifend. Dass man Morddrohungen erhält und aufgrund von Auftritten Angst um seine Familie haben muss, dürfte in einer Demokratie nicht der Fall sein.

Beschäftigt sich Parldigi auch mit Künstlicher Intelligenz? Inwiefern können internationale Abkommen wie die KI-Konvention des Europarats² als Leitlinien für ein Agendasetting der Gruppe dienen?

Min Li Marti: Im Bereich KI befinden wir uns etwas in Wartestellung, bis der Bundesrat seine versprochene Auslegeordnung in Bezug auf die KI-Regulierung und seine Botschaft zu der Plattformenregulierung³ präsentiert (Stand Juni 2024). Da es sich um eine neue Technologie handelt, könnte die KI-Konvention des Europarates durchaus als Leitlinie fungieren. Welche konkreten Auswirkungen

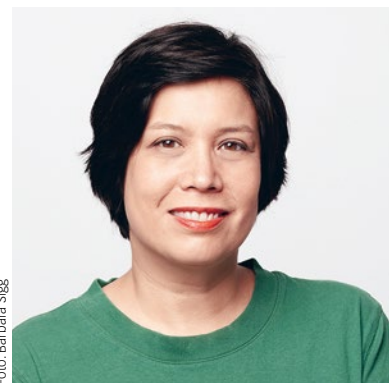


Foto: Barbara Sigg

Min Li Marti

«Weil Digitalisierung ein politisches Nischenthema ist, sind die Debatten weniger ideologisch festgefahren.»

Min Li Marti

«Die Inhalte aus den KI-Plattformen sind nicht neu und stammen oft aus den Medien.»

Franz Grüter

diese Konvention für die Schweiz hat, ist derzeit offen. Ich könnte mir aber vorstellen, dass wie beim Datenschutzgesetz gewisse Regelungen übernommen werden, das Gesetz aber nicht eins zu eins umgesetzt wird.⁴

Inwiefern ist KI im Parlament präsent und welche Technologien nutzen Sie selbst?

Min Li Marti: Da in der Kommission und der Fraktion nicht alle Dokumente von Französisch auf Deutsch übersetzt werden, nutze ich den Übersetzungsdienst DeepL⁵ häufig als Unterstützung. Zudem verwende ich eine KI-Software zur Interviewtranskription. ChatGPT⁶ habe ich ausprobiert, beschränke mich jedoch eher auf die konventionellen Suchmaschinen. Das ist vielleicht eine Generationenfrage.

Franz Grüter: Auch ich nutze DeepL oft im Parlamentsbetrieb. ChatGPT kann ich keinen grossen Nutzen abgewinnen. In meinem Umfeld merke ich, dass Schülerinnen und Schüler ChatGPT nutzen, um ihre Aufgaben schnell und einfach zu erledigen. Da die Plattformen jedoch von bereits existierenden Inhalten lernen, ist es schwierig, herauszufinden, ob die produzierten Resultate der Wahrheit entsprechen. Das ist auch deshalb herausfordernd, weil diese oft sehr eloquent präsentiert werden.

Braucht es Regulierungen in diesem Bereich?

Min Li Marti: Obwohl es bereits Software zur Plagiatskontrolle gibt, wird sich deren Einsatz in Zukunft bestimmt schwieriger gestalten, da sich die kommenden Generationen mit Verwendung von KI-Software ganze Arbeiten schreiben lassen können. Hierzu werden Schulen in Zukunft bestimmt Regulierungen erlassen müssen.

Franz Grüter: Gerne würde ich ergänzen, dass die aus den KI-Plattformen stammenden Inhalte nicht neu sind und oft aus den Medien stammen. In den USA sind bereits Klagen von Verlagshäusern gegen ChatGPT hängig. Sie machen Urheberrechtsansprüche geltend, da die von der Software gelieferten Daten aus der geistigen Arbeit ihrer Medien entstanden seien. Bestimmt wird es auch in der Schweiz bald solche Fälle geben.

Min Li Marti: Diese Problematik ergibt sich sowohl bei Chatbots wie ChatGPT als auch bei der Generierung von Bildern oder Liedern, die auf künstlicher Intelligenz beruhen. Wenn es darum geht, ein Bild

im Malstil von van Gogh oder ein Lied im Musikstil von Beyoncé zu erstellen, wird es urheberrechtlich gesehen problematisch.

Welche anderen Themen im KI-Bereich beschäftigen Sie und müssen Ihrer Meinung nach am dringendsten angegangen werden?

Min Li Marti: Mich beschäftigt die durch Algorithmen entstehende Diskriminierung sehr. Diese entsteht, weil Systeme, die mit KI funktionieren, von der Realität der Programmierenden abhängig sind. Dadurch werden andere gesellschaftliche Gruppen nicht genügend abgebildet. Algorithmen können also Diskriminierung und gesellschaftliche Ungleichheiten verstärken.

Zudem treibt mich die Frage um, ob die zunehmende Anzahl psychischer Erkrankungen bei Jungen einen Zusammenhang mit den sozialen Medien hat. Schliesslich sind die Plattformen dahingehend aufgebaut, dass ihre Nutzung süchtig macht, wodurch Isolation und Einsamkeit verstärkt werden. Sollte ein Zusammenhang bestehen, müsste meiner Meinung nach auch hier über Regulierungen diskutiert werden.

Sehen Sie bei der Vermeidung von Diskriminierung durch Algorithmen also den grössten Handlungsbedarf?

Min Li Marti: Franz Grüter und ich werden hier sicherlich nicht gleicher Meinung sein. Die Schweiz verfügt über einen relativ schwachen Diskriminierungsschutz. Im Gleichstellungsbereich ist dieser im Gesetz bedingt geregelt, hingegen sind ältere Personen oder Menschen mit einer Behinderung deutlich weniger geschützt. Meiner Meinung nach gilt es, diesem Umstand entgegenzuwirken, auch in Anbetracht der sich rasant entwickelnden Technologien.

Problematisch wird es bei Skaleneffekten. Bei der Post in England, beispielsweise, führte ein Fehler im System dazu, dass tausende Angestellte entlassen wurden. Ihnen wurden Inkorrektheiten bei der Abrechnung vorgeworfen. Nach rund zehn Jahren stellte sich jedoch heraus, dass es sich um einen Systemfehler gehandelt hatte. In den Niederlanden gab es Fälle, bei welchen aufgrund von Systemfehlern kein Kindergeld ausbezahlt wurde. Fehler und Diskriminierung werden niemals gänzlich verhindert werden, aber sie müssen korrigiert werden können.



Franz Grüter

Schlussendlich geht es auch um Akzeptanz: Inwiefern ist ein Entscheid gesellschaftlich akzeptabel, wenn er von einer Maschine anstelle eines Menschen gefällt wird? Und wie kann im Zweifelsfall Rekurs eingelegt werden? Ausserdem spielt es eine Rolle, wer die Technologie herstellt und welche Lebensrealitäten und Hintergründe diese Personen haben. Es bestehen zudem viele Ängste vor Missbrauch der Daten oder hinsichtlich der Überwachung – und das nicht zu Unrecht. Folglich ist für die Akzeptanz der Digitalisierung und neuer Technologien die Frage der Transparenz zentral. Es muss sichergestellt werden, dass Algorithmen nicht diskriminieren und nicht gegen den Gesetzesschutz agieren.

Franz Grüter: Ich lehne diese Diskussion nicht grundsätzlich ab. Für mich hat sie im gesamten Themenkomplex KI und Algorithmen jedoch einen geringeren Stellenwert als das Vorantreiben von Innovation. Ich würde es bedauern, wenn diese Frage zu einer der wichtigsten erklärt würde, während diejenigen, die uns voranbringen könnten, untergehen oder weniger stark gewichtet werden.

Also steht bei diesem Thema doch die Parteizugehörigkeit zwischen Ihnen beiden?

Franz Grüter: Nein. In unserer Gruppe werden Themen immer wieder kontrovers diskutiert und wir sind uns nicht in allem einig. Dies gehört jedoch zum Meinungsbildungsprozess und bringt uns voran.

Min Li Marti: Es ist nicht unser Anspruch, dass wir bei allen Themen gemeinsam auftreten. Unser Beitrag dient der Meinungsbildung und die verschiedenen Ansichten, die daraus resultieren, sind legitim.

Welche Mechanismen schlagen Sie vor, um gegen algorithmische Diskriminierung vorzugehen?

Min Li Marti: Da sehe ich verschiedene Möglichkeiten, wie zum Beispiel das Einsetzen einer Kommission, Qualitätskontrollen oder das Einführen einer gesetzlichen Haftung. Man könnte auch verbindlich regeln, welche Entscheide jeweils auch noch von einem Menschen kontrolliert werden müssen. Schlussendlich muss die Möglichkeit bestehen, dass illegale Aktivitäten verfolgt und ungerechte Entscheidungen angefochten werden können.

Franz Grüter: Es wäre bestimmt wichtig, dass eine Meldestelle existiert. Viele Unternehmen sind sich ihrer Fehler allenfalls gar nicht bewusst und sollten die Möglichkeit zur Korrektur erhalten. Erfolgt diese nicht, müsste die Option eines juristischen Vorgehens bestehen. In diesem Zusammenhang sollten von Seiten des Gesetzgebers Missbrauchsfelder und Korrekturbedarfe – wie zum Beispiel das Urheberrecht – identifiziert werden, bevor voreilig ein langes Gesetz verfasst wird.

Gibt es etwas, was Sie unseren Lesenden mitgeben möchten?

Min Li Marti: Ein Problem sehe ich im tiefen Diversitätsanteil in der Techbranche, sowohl in der IT als auch in anderen Bereichen. Ich erachte es als wichtig, dass sich möglichst viele verschiedene Menschen mit unterschiedlichen Hintergründen an der Technologiegestaltung beteiligen. Es ist die massgebliche Zukunftsbranche und es scheint mir entscheidend, dass auch Frauen hier mitwirken – das verbessert auch die Produkte und Lösungen. Der Frauenanteil ist aber nach wie vor tief. Vielen Frauen erscheint es schwierig, sich in einem männerdominierten Umfeld zu behaupten. Ausserdem kursieren viele Vorstellungen über Techberufe, die nicht der Realität entsprechen, beispielsweise, dass die Arbeit wenig Kreativität und Kontakt mit Menschen erlaube. Dazu kommt, dass die Berufsfindung bei Jugendlichen in einer Phase stattfindet, in der die Rollenkonformität in Bezug auf die Geschlechtsidentität und der Gruppendruck meist noch relativ gross sind.

Es gibt aber auch jenseits eines Informatikstudiums oder einer Informatiklehre Möglichkeiten, in der Techbranche zu arbeiten. Es gibt auch den Weg, per Weiterbildung oder Zweitausbildung den Einstieg zu finden, wenn vielleicht Geschlechterstereotype nicht mehr so stark wirken, denn daran scheitert es oft in jungen Jahren. Dazu kommt, dass Jobs in dieser Branche in der Regel gut mit einer Familie zu vereinbaren sind.

Franz Grüter: Damit stimme ich überein. Ich komme selbst aus der IT-Industrie, in welcher 20 Prozent Frauen und 80 Prozent Männer tätig sind. Das verändert sich bedauerlicherweise auch beim Nachwuchs kaum. Es wäre spannend zu ergründen, wieso in meinem Empfinden viele Frauen in der Schweiz eine solch starke Abneigung haben, sich in

«Ich erachte es als wichtig, dass sich möglichst viele verschiedene Menschen mit unterschiedlichen Hintergründen an der Technologiegestaltung beteiligen.»

Min Li Marti

«Es wäre spannend zu ergründen, wieso in meinem Empfinden viele Frauen in der Schweiz eine solche starke Abneigung haben, sich in der IT-Industrie zu engagieren.»

Franz Grüter

der IT-Industrie zu engagieren. Dieses Verhältnis ist schliesslich in anderen Ländern in Asien, Südamerika und in den USA ganz anders. Ich wünschte mir, dass hier eine Verbesserung herbeigeführt wird.

Nadja Senn ist wissenschaftliche Mitarbeiterin bei der Kantonalen Opferhilfestelle Zürich und war von 2022 bis 2023 Hochschulpraktikantin bei der EKF.

Anmerkungen

- 1 www.parldigi.ch
- 2 Europarat: Europarat verabschiedet ersten internationalen Vertrag über Künstliche Intelligenz, 2024. <https://www.coe.int/de/web/portal/-/council-of-europe-adopts-first-international-treaty-on-artificial-intelligence> (abgerufen am: 22.05.2024).
- 3 Der Bundesrat: Grosse Kommunikationsplattformen: Bundesrat strebt Regulierung an. [Medienmitteilung], 05.04.2023. <https://www.admin.ch/gov/de/start/dokumentation/medienmitteilungen.msg-id-94116.html>.
- 4 Eidgenössisches Justiz- und Polizeidepartement EJPD: Künstliche Intelligenz und Recht. [Rede], 22.03.2024. Konferenz zur Regulierung Künstlicher Intelligenz in Zürich, 2024. <https://www.admin.ch/gov/de/start/dokumentation/reden/reden-der-bundesraete.msg-id-100501.html#:~:text=Der%20Bundesrat%20hat%20aber%20das,EU%20und%20im%20Europarat%20ber%C3%BCcksichtigen.>
- 5 www.deepl.com
- 6 <https://chatgpt.com>

Abstracts

Où en est la politique suisse dans le domaine de l'IA ? La coprésidence de Parldigi répond

La conseillère nationale PS **Min Li Marti** et le conseiller national UDC **Franz Grüter** se partagent la présidence de Parldigi. Il s'agit d'un groupe parlementaire qui étudie les questions de transformation numérique, dépose des interventions et organise des manifestations de sensibilisation. Dans leur interview, les deux parlementaires portent une appréciation sur les opportunités et les risques que présentent les outils d'IA et ils discutent de la nécessité d'une réglementation en Suisse. Ils montrent que, malgré leurs avis souvent divergents sur l'IA et la discrimination, ils arrivent régulièrement à trouver des consensus qui dépassent les clivages partisans.

Qual è la posizione della politica svizzera in materia di IA? Rispondono i due co-presidenti di Parldigi

La consigliera nazionale del PS **Min Li Marti** e il consigliere nazionale dell'UDC **Franz Grüter** co-presiedono Parldigi. Questo gruppo parlamentare si occupa di questioni legate alla digitalizzazione, presenta interventi parlamentari e svolge attività di sensibilizzazione attraverso manifestazioni, valutano le opportunità e i rischi degli strumenti basati sull'intelligenza artificiale. In quest'intervista, discutono del bisogno di regolamentazione in Svizzera e mostrano come, sebbene spesso le loro opinioni sul tema dell'intelligenza artificiale divergano, riescano sempre a trovare consensi interpartitici.

Feline Surveillance

Hi DALL-E, can you please produce an image in portrait format in photo realistic digital art style depicting a very cute cat on a couch? The cat is actually a robot disguised as a cat. The animal is analyzing everything we do through its eyes and getting a lot of information.

We are experiencing heavy server load. To ensure the best experience for everyone, we have rate limits in place. Please wait for 2 minutes before generating more images. If you have any other requests or need assistance, feel free to let me know!

Can we try again?

Here is a photorealistic image of a fluffy grey cat sitting in a tree. The cat is actually a robot disguised as a cat, with subtle mechanical elements visible under its fur, especially around its eyes which are high-tech lenses. The cat is analyzing everything around it through its eyes, gathering a lot of information. The tree and background are realistic, highlighting the cat as the main focus. There are also ants crawling on the tree, adding more detail to the scene.



Zeitschrift **Frauenfragen**

46. Jahrgang, Ausgabe November 2024

Revue **Questions au féminin**

46^e année, édition novembre 2024

Rivista **Questioni femminili**

46° anno, edizione novembre 2024

Herausgeberin / Édition / A cura della

Eidgenössische Kommission für Frauenfragen EKF / Commission fédérale pour les questions

féminines CFQF / Commissione federale per le questioni femminili CFQF

Schwarztorstrasse 51, 3003 Bern, 058 462 92 75

ekf@ebg.admin.ch, www.frauenkommission.ch, www.comfem.ch

Konzept / Concept / Concetto

Caroline Honegger, Bettina Fredrich

Redaktion / Rédaction / Redazione

Caroline Honegger, Bettina Fredrich, Mathilda Golay

Verantwortlich / Responsable / Responsabile

Caroline Honegger

Übersetzung / Traduction / Traduzione

Catherine Kugler, Sandra Verzasconi Catalano, IZE Language Services

Abonnemente / Abonnements / Abbonamenti

«**Frauenfragen**» erscheint einmal pro Jahr. Das Abonnement ist kostenlos und kann bestellt werden bei: Bundesamt für Bauten und Logistik BBL, Bundespublikationen, CH-3003 Bern, www.bundespublikationen.admin.ch

www.frauenkommission.ch > Publikationen > Fachzeitschrift «Frauenfragen»

«**Questions au féminin**» paraît une fois par an. L'abonnement est gratuit et peut être obtenu auprès de: Office fédéral des constructions et de la logistique OFCL, Publications fédérales, CH-3003 Berne, www.publicationsfederales.admin.ch

www.comfem.ch > Publications > Revue spécialisée «Questions au féminin»

«**Questioni femminili**» esce una volta all'anno. L'abbonamento è gratuito e può essere richiesto presso: Ufficio federale delle costruzioni et della logistica UFCL, Pubblicazioni federali, CH-3003 Berna, www.pubblicazionifederali.admin.ch
www.comfem.ch > Pubblicazioni > Rivista specializzata «Questioni femminili»

Gestaltung / Graphisme / Concezione grafica

Renata Hubschmied, Bern

Titelbild / Image de couverture / Immagine di copertina

Alizé Rose-May Monod, DALL-E

Die in den einzelnen Artikeln vertretenen Standpunkte müssen sich nicht mit der Meinung der EKF decken.

Les points de vue exprimés dans les divers articles ne doivent pas forcément coïncider avec l'opinion de la CFQF.

I punti di vista espressi nei diversi articoli non devono necessariamente corrispondere a l'opinione della CFQF.

ISSN: 2235-1833, Art. Nr. 303.500.24 11.24 4500 860 557 292

Künstliche Intelligenz und algorithmische Systeme erleichtern unseren Alltag massgeblich. Sie lernen täglich dazu und liefern präzisere Resultate. Dabei stützen sie sich auf vorhandene Daten. Gerade aus Gleichstellungsperspektive ist das nicht unproblematisch, denn häufig reproduzieren die Daten gesellschaftliche Ungleichheiten. Im vorliegenden Heft fragen wir, welche technischen, politischen, rechtlichen und gesellschaftlichen Anpassungen nötig sind, damit die neue Technologie die Geschlechtergleichstellung verbessert. Wo liegen die Chancen und was sind die Herausforderungen? Mit Fachartikeln, Interviews und Porträts beleuchten wir die aktuelle Situation in der Schweiz.

L'intelligence artificielle et les systèmes algorithmiques facilitent considérablement notre quotidien. Ils accroissent leurs connaissances de jour en jour et fournissent des résultats de plus en plus précis. Mais ils s'appuient sur des données existantes, ce qui pose problème du point de vue de l'égalité car, souvent, les données reproduisent les inégalités sociales. Dans la présente édition de notre revue, nous nous demandons quelles sont les adaptations techniques, politiques, juridiques et sociales qui sont nécessaires pour que cette nouvelle technologie améliore l'égalité des genres. Quelles opportunités et quels défis recèle-t-elle ? Des articles, des interviews et des portraits éclairent la situation actuelle en Suisse.

L'intelligenza artificiale e i sistemi algoritmici ci semplificano notevolmente la vita quotidiana. Ogni giorno imparano cose nuove e forniscono risultati più precisi. A tal fine, si basano sui dati disponibili. Ciò non è scevro di problemi, soprattutto dal punto di vista dell'uguaglianza, poiché spesso tali dati riproducono disuguaglianze sociali. In questo numero di «Questioni femminili» ci domandiamo quali adeguamenti tecnici, politici, giuridici e sociali siano necessari affinché la nuova tecnologia migliori l'uguaglianza di genere. Quali opportunità offre e quali sfide pone? Con articoli specialistici, interviste e ritratti facciamo luce sulla situazione attuale in Svizzera.