SocietyByte

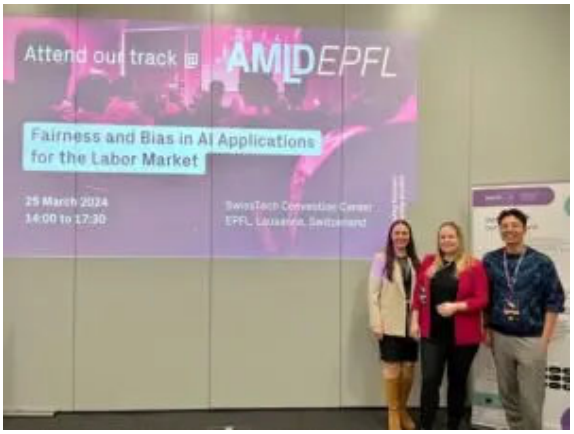
BFH-Magazin für die Humane Digitale Transformation

Fairness und Voreingenommenheit bei KI-Anwendungen für den Arbeitsmarkt

Von Elena Nazarenko , Alexandre Puttick (BFH Technik & Informatik) | 0 Kommentare

Für die Konferenz Applied Machine Learning Days (AMLD) 2024 an der EPFL organisierten die BFH-Gruppe Applied Machine Intelligence [https://www.bfh.ch/en/research/research-areas/applied-machine-intelligence/] und die NLP-Expertin Dr. Elena Nazarenko einen Track über Fairness bei KI-Anwendungen auf dem Arbeitsmarkt. Die Konferenz bringt über 1500 Teilnehmer aus über 40 Ländern aus Industrie, Wissenschaft und Regierung zusammen.

Für die diesjährige Ausgabe der Konferenz organisierte die Forschungsgruppe Angewandte Maschinelle Intelligenz der Berner Fachhochschule [https://www.bfh.ch/de/forschung/forschungsbereiche/applied-machine-intelligence/] zusammen mit Dr. Elena Nazarenko von der Hochschule Luzern einen Track zum Thema Fairness bei KI-Anwendungen auf dem Arbeitsmarkt. Der Track brachte akademische und industrielle Perspektiven zusammen und stützte sich auf ein breites Spektrum von Disziplinen, darunter Datenwissenschaft, Recht, Philosophie, Wirtschaft und Psychologie. Nach einer kurzen Einführung durch die Mitorganisatorinnen des Tracks **, Elena Nazarenko** und **Mascha Kurpicz-Briki**, stellte Mascha **Kurpicz-Briki** BIAS vor, ein von der EU und der Schweiz finanziertes Projekt zur Verringerung von KI-Voreingenommenheit auf [https://www.bfh.ch/en/research/research-projects/2022-025-172-803/] dem Arbeitsmarkt, für das sie eine der technischen Leiterinnen ist.



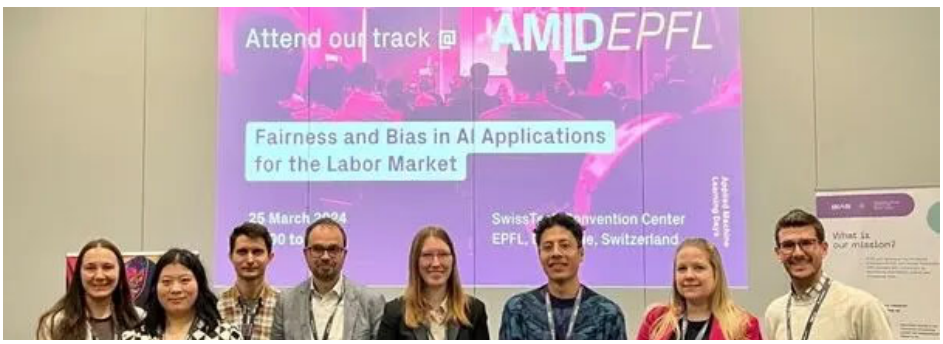
Prof. Dr. Mascha Kurpicz-Briki (in der Mitte) und die Autoren dieses Artikels vor ihrem Bildschirm.

Eduard Fosch-Villaronga, ausserordentlicher Professor an der Universität Leiden und Leiter des auf Recht spezialisierten Teams im Rahmen des BIAS-Projekts, hielt den ersten Vortrag. Er beschrieb die Reibungen zwischen dem Fairnessverständnis von Bewerbern und Personalverantwortlichen und stellte dieses Verständnis in den Kontext des EU-Rechts und des AI-Gesetzes. Seine Forschung zeigt auch die Bedeutung des Themas, indem er feststellt, dass KI-Anwendungen auf dem Arbeitsmarkt bereits weit verbreitet sind. Schliesslich beschrieb er die Positionen von KI-Entwicklern, HR-Praktikern und der allgemeinen Bevölkerung zum Einsatz von KI auf dem Arbeitsmarkt. Es gibt eine weit verbreitete Anerkennung der potenziellen Vorteile und Nachteile, aber auch Unsicherheit darüber, wann und wie solche Werkzeuge eingesetzt werden können/sollten.

Der nächste Vortrag wurde von **Preethi Lahoti**, Research Scientist bei Google, gehalten, die den Prozess der Erstellung von sicheren, integrativen und fairen Large Language Models (LLMs) erörterte. Sie sprach über die Verwendung von LLMs, um ihre eigene Sicherheit und Fairness zu verbessern, und konzentrierte sich dabei auf gegnerische Tests und Minderungsstrategien. Sie stellte AI-assisted Red Teaming (AART) für gegnerische Tests vor, bei denen ein LLM seine eigenen Prompts generiert, um nach schädlichen Antworten zu suchen, sowie eine neuartige Methode namens Collective-Critiques and Self-Voting (CCSV), die die Vielfalt in der Ausgabe eines LLMs verbessert, indem es mehrere Antworten für einen bestimmten Prompt generiert, kritisiert, verbessert und darüber abstimmt.

Als letzter Redner der ersten Sitzung befasste sich **Alejandro Jesús Castañeira Rodriguez** von Janzz.technology mit der kritischen Frage der Fairness und Voreingenommenheit bei KI-gestützten Empfehlungen innerhalb der Belegschaft. Er stellte das Empfehlungssystem von Janzz Technology vor. Das System unterscheidet sich von reinen maschinellen Lernsystemen dadurch, dass es sich ausdrücklich auf relevante Merkmale und Wissensgraphen stützt, wodurch es die Transparenz erhöht und unfaire Verzerrungen abschwächt.

Die Nachmittagssitzung begann mit einem Vortrag von **Christoph Heitz**, Professor an der ZHAW, der sich auf algorithmische Fairness spezialisiert hat. Er stellte verschiedene Standpunkte zur Fairness vor, die von Aristoteles bis zum *Techno-Solutionismus* reichen, dem Glauben, dass Technologie alle sozialen Probleme lösen kann. Er argumentierte, dass wir zur Schaffung eines fairen algorithmischen Systems zunächst eine ethische Haltung definieren müssen. Anschließend benötigen wir entsprechende Fairness-Metriken, die den Grad der Übereinstimmung des Systems mit der gewählten Vorstellung von Fairness messen sollen. Sein Vortrag schloss mit einem vereinheitlichenden Rahmen für bestehende Fairness-Metriken, der zeigt, dass jede von ihnen aus einer spezifischen Wahl der Nutzenfunktion, der demografischen Gruppen und der *Rechtfertiger*, d.h. der moralischen Gründe für gerechtfertigte Ungleichheit, resultiert.



Elena Nazarenko (links), Alexandre Puttik (3. von rechts) und Mascha Kurpicz-Briki (2. von rechts) mit Kollegen

Die nächste Referentin war **Cynthia Liem**, eine ausserordentliche Professorin an der TU Delft. Sie beschrieb die fächerübergreifende Arbeit mit Psychologen an einem algorithmischen Bewerberscreening sowie eine in Kürze erscheinende Arbeit über mathematische Begriffe der Fairness. Sie wies auf wichtige Missverständnisse zwischen den Disziplinen hin. So wurden beispielsweise Persönlichkeitstests sehr fragwürdig eingesetzt, um Daten für die automatische Auswertung von Videointerviews zu kennzeichnen. Darüber hinaus zeigen ihre jüngsten Forschungen, dass kein einziger mathematischer Fairness-Begriff für eine frühzeitige Bewerberauswahl geeignet ist.

Im Anschluss diskutierte **Jana Mareckova**, Assistenzprofessorin für Ökonometrie an der Universität St. Gallen, den Einsatz von kausalem maschinellem Lernen zur Bewertung der Wirksamkeit von Arbeitslosenprogrammen wie Kursen und Subventionen. Ihr Vortrag behandelte die Messung von Behandlungseffekten für verschiedene Gruppen sowie für eine bestimmte Person. Letzteres stellt eine besondere Herausforderung dar, da wir in der Praxis nicht beobachten können, was passiert wäre, wenn die Person ein anderes Programm oder gar kein Programm absolviert hätte. Sie hob die Einsichten hervor, die diese Modelle bieten, und die Herausforderungen, die sie bei der Interpretation mit sich bringen.

Am Ende der Sitzung präsentierte **Pencho Yordanov**, Lead Data Scientist bei der Adecco Group, seine Arbeit über den Einsatz von LLMs in Hochrisikosektoren wie dem Personalwesen. Er hob die Rolle der Psychologie beim Verstehen und Abschwächen von Verzerrungen bei der Entscheidungsfindung hervor und beschrieb, wie menschliche kognitive Verzerrungen auch in LLMs vorhanden sind und wie solche Verzerrungen die Bewerberauswahl beeinflussen. Er erläuterte den Einfluss von *Scheinkandidaten*, die sehr ähnlich, aber etwas schlechter als andere Kandidaten sind, auf die Präferenzen von Personalverantwortlichen und beobachtete ähnliche Effekte in Experimenten zu GPT3.5 und GPT4, was eine weitere Ebene der Vorsicht aufzeigt, die beim Einsatz solcher Systeme beachtet werden muss.

Über AMLD

Die Applied Machine Learning Days finden an der EPFL in Lausanne statt. Es handelt sich um eine der grössten Veranstaltungen für maschinelles Lernen und KI in Europa, die sich speziell auf die Anwendungen von maschinellem Lernen und KI konzentriert. Die Veranstaltung bringt mehr als 2.000 Führungskräfte, Experten und Enthusiasten aus Wissenschaft, Industrie, Start-ups, NGOs und Behörden aus mehr als 41 Ländern zusammen.

AUTHOR: ELENA NAZARENKO

Elena Nazanrenko ist Datenwissenschaftlerin bei der Zürcher Entwicklerfirma Witty Works. Sie hat einen Hintergrund in theoretischer und rechnergestützter Physik und entwickelte u.a. bereits ein NLP-Projekt für kollaboratives Arbeitsmanagement, einen Chatbot-Prototyp und verbessere die Freitextsuche einer eCommerce-Plattform. Zuvor arbeitete sie als Wissenschaftlerin am Paul Scherrer Institut (ETH-Bereich, Schweiz) und an nationalen Forschungsinstituten in Schweden und Frankreich.

Posts from Elena Nazarenko | Website

AUTHOR: ALEXANDRE PUTTICK

Dr. Alexandre Puttick ist Post-Doktorand in der Forschungsgruppe Angewandte Maschinelle Intelligenz an der Berner Fachhochschule. Seine aktuelle Forschung befasst sich mit der Entwicklung von klinischen Tools für die psychische Gesundheit sowie mit der Erkennung und Abschwächung von Verzerrungen in KI-gesteuerten Rekrutierungs-Tools.

Posts from Alexandre Puttick

Create PDF

Ähnliche Beiträge

Applied NLP Technologies for Physical and Mental Health

0

COMMENTS