fAIr by design

fAIr by design – solutions for discrimination reduction in AI development

Programm / Ausschreibung	Laura Bassi 4.0, Laura Bassi 4.0, Laura Bassi OEF 2018	**Status**	abgeschlossen
Projektstart	01.04.2021	**Projektende**	31.03.2024
Zeitraum	2021 - 2024	**Projektlaufzeit**	36 Monate
Keywords	AI, Fairness, Algorithmen, Diskriminierung, Gender, Alter, Ethnizität		

Projektbeschreibung

Artificial Intelligence (AI) beeinflusst nachhaltig das Leben von Menschen. Vor allem in kritischen Anwendungsfeldern (Arbeit – HR, Health, etc.) ist in naher Zukunft mit erhöhten Anforderungen in Bezug auf Ethik, Rechtskonformität und technischer Robustheit zu rechnen, um zu garantieren, dass AI-Anwendungen nicht zu negativen gesellschaftlichen Wirkungen führen und in Summe vertrauenswürdig sind (Trustworthy AI Framework). Staatliche Regulierung und freiwillige Zertifizierungserfordernisse werden künftig den Markt prägen und den Druck auf Unternehmen erhöhen, AI-Anwendungen zu entwickeln, die hohen ethischen Standards genügen. Dazu gehört auch der Anspruch, nicht zu diskriminieren.
Eine der größten ungelösten Herausforderungen liegt in der Frage, wie die Diskriminierung von Nutzergruppen durch AI bereits während der Entwicklung bzw. vor der Anwendung von AI im Markt vermieden werden kann. Bei der Vermeidung von diskriminierendem Bias müssen sozialwissenschaftliche, rechtliche, ethische und technologische Fragen adressiert werden, wie sie bisher in der AI-Entwicklung in dieser Interdisziplinarität nur wenig Beachtung gefunden haben und in den herkömmlichen Prozessen und Prüfverfahren der AI-Entwicklung nicht abgebildet werden. Somit verfügen AI-Entwickler*innen und Anwender*innen (z.B. Unternehmen, öffentliche Hand) derzeit über unzureichend anwendungs-orientierte Werkzeuge, die den gesamten AI-Entwicklungsprozess abbilden, um die Diskriminierung von bestimmten Nutzergruppen durch AI-Anwendungen frühzeitig vermeiden, erkennen und verhindern zu können.
Dieses Manko adressiert das Projekt fAIr by design, welches unter Einbeziehung von unterschiedlichen Nutzergruppen und 5 Use Cases auf die Entwicklung eines neuartigen Verfahrensmodells und einer entsprechenden Methodentoolbox für die Entwicklung von fairer, nicht-diskriminierender AI abzielt. Unter Anwendung von Open Innovation Methoden entwickeln Data Scientists, AI-Expert*innen, Sozialwissenschaftler*innen, Rechtsexpert*innen und die jeweiligen Anwender*innen aus Unternehmen und anderen Organisationen Module und Strategien zur Risikoreduzierung für unterschiedliche Diskriminierungsrisiken, welche danach in einem generischen Prozess und einer Methodentoolbox für die Diskriminierungs-Vermeidung zusammengefasst werden. Da-bei steht ein Spektrum von AI mit direkter Auswirkung auf Menschen in Gesellschaft und Arbeit, z.B. im Bereich Ausbildung, HR, Medien und Health durch die bearbeiteten Use Cases im Vordergrund, insgesamt soll jedoch ein Prozess und eine Toolbox entstehen, die möglichst breit in allen möglichen Diversitäts-/Diskriminierungsdimensionen anwendbar sein soll.
Das Konsortium besteht aus 8 Organisationen, welche relevantes Vorwissen inkl. aktueller Use Cases einbringen, die

Entwicklung von fairer AI voranbringen und einen direkten Benefit aus der Verwertung der Projektergebnisse ziehen können, davon vier KMU, zwei Universitäten, eine gemeinnützige GmbH und ein Großunternehmen.

Abstract

Today Artificial Intelligence (AI) has a lasting effect on people's lives. Particularly in critical applica-tion areas (working life - human resources, health, etc.), increased requirements in terms of ethics, legal conformity and technical robustness can be expected in the near future in order to guarantee that AI applications do not lead to negative social effects and are trustworthy overall (Trustworthy AI Framework). Both government regulation and voluntary certification requirements will shape the AI market in the future, increasing the pressure on companies to develop AI applications that do not discriminate or discriminate only to an extent that is unavoidable and is disclosed in a transparent manner.
One of the biggest unresolved challenges is how to avoid discrimination of user groups by AI already during the development phase, however before AI deployment in society and markets. In order to avoid socially and technologically discriminating bias, ethical considerations and social science methods have to be incorporated, which is not considered in the AI development up to now and not reflected in the conventional processes and test procedures of AI development. Thus, AI developers and AI users (e.g., companies, public institutions) currently have insufficient applica-tion-oriented tools that cover the entire AI development process in order to avoid, recognize, and prevent undesired discrimination of certain user groups by AI applications at an early stage.
This shortcoming is addressed by the project fAIr by design, which aims at the development of a novel generic procedural model and a corresponding method toolbox for the development of fair, non-discriminatory AI, involving different user groups and 5 use cases. Using open innovation methods, data scientists, AI experts, social scientists, legal experts and the respective application experts from companies and other organizations will develop modules and strategies for risk reduction for different discrimination risks, which will then be concluded and further developed in a generic process model and a method toolbox for the prevention of discrimination. The focus is on a spectrum of AI with a direct impact on people in society and work, e.g. in the areas of education, human resources, performance assessment, media and health, through the use cases worked on, but overall a process and a toolbox should be created that can be applied as broadly as possible in all possible diversity/discrimination dimensions.
The consortium consists of 8 organizations, which contribute relevant prior knowledge including current use cases, promote the development of fair AI and can draw a direct benefit from the exploitation of the project results, including four SMEs, two universities, one non-profit limited company and one large enterprise.

Projektkoordinator

- winnovation consulting gmbh

Projektpartner

- Cultural Broadcasting Archive, Verein zur Förderung digitaler Kommunikation
- rotable technologies GmbH
- Speedinvest Heroes Consulting GmbH
- Intact GmbH
- Universität Wien
- Technische Universität Wien
- Rania Wazir e.U.