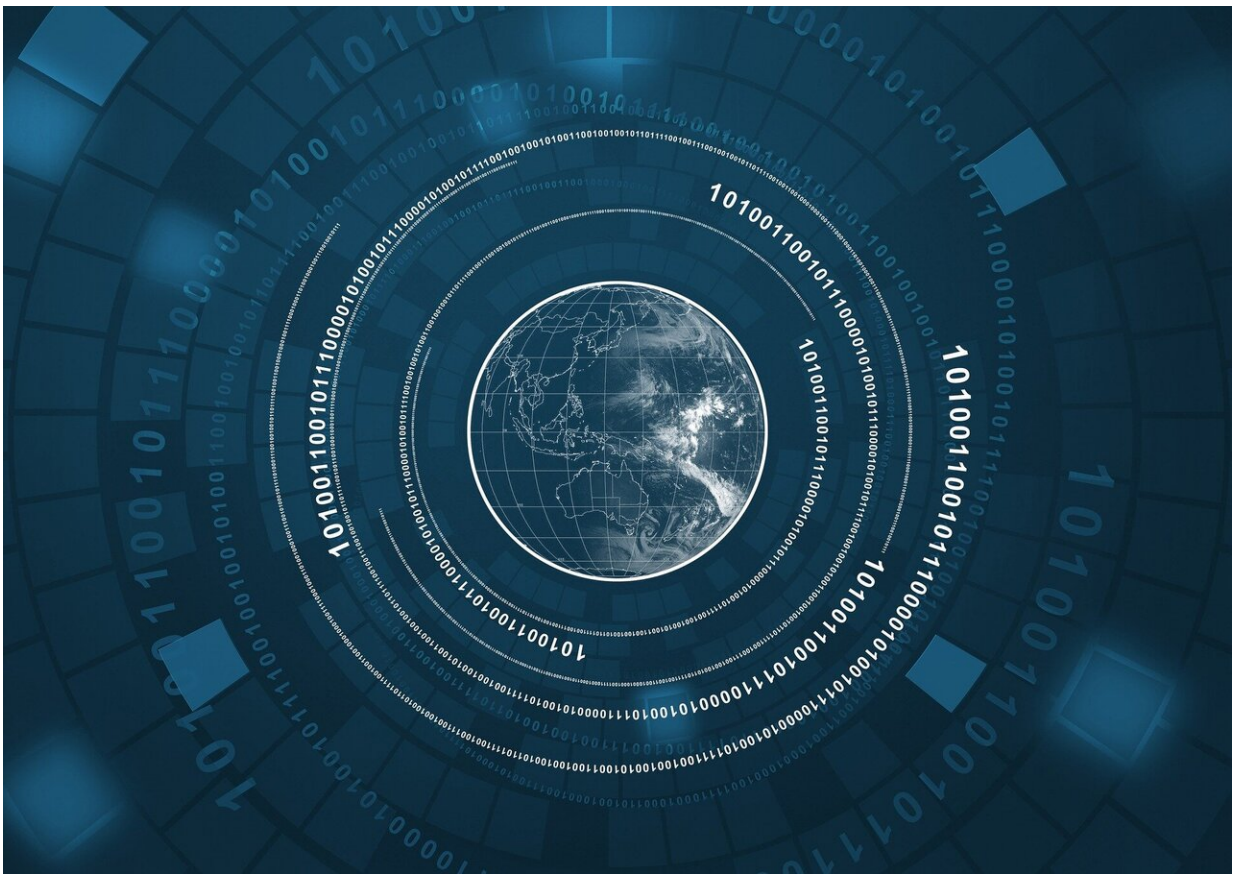
The Achilles' heel of artificial intelligence: Why discrimination remains an unresolved problem

January 10 2024, by Carolin Höll



Credit: CC0 Public Domain

A recent study by the DHBW Stuttgart at the Service Management Study

Center (ZMM) investigates the ability of Artificial Intelligence (AI) to recognize discriminatory content in images and advertisements, showing both impressive progress as well as existing limitations.

In this comprehensive study, AI was confronted with a variety of images and advertisements and asked to evaluate them. This included 60 advertisements, among others, that the German Advertising Council had recently criticized. The results show that AI has an astonishing ability to identify discriminations in advertisements with impressive accuracy (F1 Score: 0.949). As a result, the AI generally assessed the advertisements criticized by the German Advertising Council as potentially discriminatory. At the same time, in most cases, non-discriminatory advertisements did not receive such a warning.

This is particularly impressive considering that just 10 years ago, AI had significant difficulties in correctly classifying objects depicted in an image. In the meantime, and thanks to millions of images, AI has learned to largely accurately distinguish between a dog and a cat. The rapid progress of AI raises the question of where its current limits lie. "We wanted to know to what extent AI recognizes discriminatory behavior when it is only presented with an advertisement and asked to evaluate it," explains student Helen Beckers about the approach of the study.

This is of particular relevance in light of the massive increase in discrimination by algorithms that can disadvantage people based on gender, religion, ideology, racism, or origin. This aspect is becoming more important as discrimination by algorithms is increasingly becoming a serious problem affecting various areas such as application processes, credit allocation, medicine, and the calculation of the recidivism probability of offenders.

The insight that ChatGPT can detect sexualization and stereotypical thinking is particularly revealing. A modified advertisement with

swapped gender roles showed that AI can also identify discrimination in reversed situations. "This different evaluation by ChatGPT in the two scenarios highlights the ability to identify discrimination even in reversed situations," says student Sven Peter, sharing another finding from the study.

However, AI reached its limits in identifying other forms of discrimination, such as objectification, disrespect, and abuse of power. "The results of the study underscore the need to further develop AI systems to recognize discrimination more effectively and prevent it," reports student Marius Funk. The study raises important questions about how AI technologies can be used in the future to combat discrimination in various areas and promote equality.

Provided by Duale Hochschule Baden-Württemberg

Citation: The Achilles' heel of artificial intelligence: Why discrimination remains an unresolved problem (2024, January 10) retrieved 20 May 2024 from https://techxplore.com/news/2024-01-achilles-heel-artificial-intelligence-discrimination.html