Addressing Discrimination in Artificial Intelligence

Federica Picogna, PhD student at Nyenrode Business Universiteit
Jacques de Swart, Professor of Applied Mathematics at Nyenrode Business Universiteit
Ruud Wetzels, Professor of Data Science at Nyenrode Business Universiteit

In recent years there has been a rapid development of Artificial Intelligence (AI) and an increase in its use by humans. More specifically, there has been a focus on its ability to assist and even replace humans in the decision-making process. Although this creates opportunities for increasing efficiency and transparency in decision making, it can also be risky for both citizens, businesses, as well as the government.

For example, the Netherlands Enterprise Agency (RVO) uses an algorithm to facilitate the assessment of grant applications submitted for the government's *Tegemoetkoming Vaste Lasten* (TVL) scheme. This program aims to reimburse business owners for fixed costs incurred due to the government's measures to combat the COVID-19 pandemic. The use of this algorithm allows for an efficient processing of grant applications, ensuring quick disbursement of funds for applications classified as low-risk.
However, the use of the algorithm also poses the risk of public funds being misused or improperly allocated when a high-risk application is mistakenly categorized as a low-risk one. This could have negative repercussions for the government. Simultaneously, if a low-risk application is mistakenly classified as a high-risk one, manual scrutiny of the application is required, causing delays in fund distribution and additional losses for business owners.

The increasing awareness of the risks associated with the use of AI system, as well as the need to mitigate these, has led to a demand for accountability in its usage. It is also the auditor's responsibility to identify any improper use of AI in the decision-making process, following benchmark regulations such as the European AI Act.

The auditor, therefore, will engage in Responsible AI. This field involves the application of a series of techniques for a better understanding of the decision-making process and its output. In this way, it is possible to enhance transparency and ensure that the output does not lead to discriminatory actions. Discrimination, in this context, is a sociological term that refers to the unfair and unequal treatment of individuals in a certain group based solely on their affiliation with that group, depriving the members of one group of benefits accessible to other groups. For example, in the case of an algorithm used by the RVO or other agencies that grant applications, it can be possible that there is discrimination when applications with a history of immigration are more frequently classified as high-risk compared to those without such a history.

In the context of Responsible AI, a machine learning algorithm for classification needs to simultaneously optimize utility for a specific purpose while preventing discrimination against protected population subgroups. To date, there are various statistical measures to assess discrimination in the outcome of an algorithm. These measures, used by the auditor, are known as fairness measures. For example, in the case of the algorithm used by the RVO, an auditor might consider verifying the absence of discrimination by comparing the rates of applications correctly classified as high-risk for those with and without an immigration history.

The presence of various fairness measures with the same overall objective is linked to the fact that, depending on the situation, certain measures may be more suitable than others, given their distinct properties. However, the hasty

development of these measures has had some negative repercussions. In fact, numerous measures, despite having different names, are essentially the same, adding further difficulty when the auditor must choose among them. Furthermore, auditors may not possess an advanced statistical education. Coupled with the constant evolution of these statistical tools and the absence of consensus in defining discrimination, bias, and fairness, significant confusion arises. This confusion can result in drawing inaccurate conclusions about the existence or absence of discrimination, potentially leading to severe consequences.

Hence, there is a demand for innovation in this field, not only to address the aforementioned challenges but also to correctly guide auditors in determining the presence or absence of discrimination.
From these needs arises the doctoral project, promoted by Nyenrode Business Universiteit and in which we are involved, with the goal of developing advanced Bayesian statistical methods within the responsible AI domain. Subsequently, these methods will be easily accessible to auditors through the open-source program JASP that will enable the simplest approach to the problem, eliminating the technical programming difficulty associated with it.

The main goal is to develop a decision workflow that can guide auditors in selecting the most suitable fairness measures based on the context under consideration. To help users choose the best fairness measures, we suggest using interactive questions that guide them to a subset of measures tailored to their needs. This is particularly necessary since fairness measures can be mutually exclusive, highly context-dependent, and quite complex. When asking these questions, we follow the principle of transitioning from general inquiries to progressively more specific ones.

This workflow is intended to be used for audit purposes, especially by those without a strong statistical background. However, determining the most suitable measure for a specific case study requires comprehending the properties and aspects that are important to capture and adhere to, particularly those respected by the statistical characteristics of the measures. In this context, a comparison between various scenarios becomes essential, as does the exchange of ideas and collaboration across diverse knowledge domains.

Noten

[1] Algemene Rekenkamer, "An audit of algorithms: Nine algorithms used by the Dutch government," May 2022.