WHITEPAPER

Certifying Fairness of AI-Applications An Impossible Task?

Sebastian Scher and Sarah Stryeck
January 2022

Den Rise/shutterstock.com

INTRODUCTION

Since more and more decisions that used to be made by humans are nowadays made either with the help of Artificial Intelligence (AI) or by AI alone, it is essential that the AI algorithms are "fair." In this whitepaper, we discuss issues surrounding the fairness of AI applications, with a special focus on how it can be assessed independently and subsequently certified. We explain why an AI application cannot be classified – and subsequently certified – as "fair" or "unfair" in a general sense and propose an approach that makes it possible to classify it as "fair" or "unfair" under the (application)-specific definition of fairness.

CONTENTS

MISSION IMPOSSIBLE?

As Artificial Intelligence (AI) is becoming increasingly popular, concerns about its application are rising. One of the most controversial issues is that AI-based applications could result in an unfair outcome, e.g., by favoring men over women or people of a specific ethnicity over people of other ethnicities. Issues like that can be addressed by establishing ways of determining whether an AI-based application is "fair"[1,2]. However, this is a very difficult task since "fairness" is difficult to define.

Most people have a good sense of fairness and - even more so - unfairness when evaluating existing outcomes. For example, if an algorithm used for job applications consistently rejects applications by women, most people would call it unfair. However, the question "What is fair?" is much harder to answer, and people are typically less willing to provide quick and precise responses to it. One potential response in our example could be "The algorithm should pick an equal number of men and women." Intuitively, this sounds fair. But what if there are significantly more male applicants than female ones? 10 out of 10 women and 10 out of 50 men would appear unfair to most people. It may be tempting to make it right by taking the same fraction of each, i.e., 2 out of 10 women and 10 out of 50 men. But what if there are big differences in the educational backgrounds among female and male applicants? What would a fair outcome be in that case? Another intuitive answer could be "Men and women should be treated equally." But how can equality be verified? And wouldn't this be contrary to the intuitive fairness idea that a roughly equal number of men and women should be considered? Even this simple example shows how challenging it is to find objective ways of certifying something as "fair."

Another fundamental issue is that such concepts as fairness are deeply rooted in cultural beliefs, i.e., what is considered fair or unfair varies significantly between different countries, regions, ethnic groups, age groups, etc. In addition, the notion of fairness can - and most probably will – evolve and change over time. All this makes certifying something as 'fair' independently and objectively almost impossible.

Nevertheless, there clearly is a demand for ascertaining the fairness of applications since fairness is closely linked to non-discrimination, which is a fundamental right according to Article 21 of EU's[3] Charter of Fundamental Rights. Non-discrimination also plays an important role in the recently proposed EU AI regulation, and the EU's Fundamental Right Agency (FRA) has published several reports on algorithmic fairness and related issues[4,5,6]. For businesses, preventing legal problems and public setback is an important incentive for ensuring that their AI applications are fair and non-discriminatory.

In this paper, we provide an overview of ways to certify the fairness aspects of AI applications and demonstrate how the problem of objective fairness certification can be solved.

fizkes/shutterstock.com

DEFINING FAIRNESS

In the introduction, we have mentioned some fundamental issues with defining fairness (e.g., different intuitive fairness definitions that contradict each other and the general idea of fairness being significantly affected by culture).

Based on the realization that different concepts of fairness may be contradictory, two main concepts of fairness have been established: group fairness and individual fairness[7].

Group fairness means that groups as a whole are treated equally. In our example of job application algorithm described in the introduction, this means that the group of women as a whole is treated equally to the group of men. In this case, "equal treatment" could mean that an equal fraction of both is accepted or that, on average, the quality of predictions is the same for both groups (e.g., the fraction of excellent candidates erroneously rejected by the algorithm).

Individual fairness means that individuals are considered treated equally if they are equal regardless of the attributes (e.g., gender, ethnicity). The meaning of "equal individuals" depends on the context and the application. In our example of job application algorithm, individual fairness means that individuals with equal attributes relevant for the job (e.g., education, experience, grades) are treated equally and independently of their gender, etc.

Individual fairness, which is closely related to how non-discrimination is legally defined, sounds very intuitive at first. However, there are lots of issues associated with it. The available attributes (e.g., experience in our example) cannot always be assessed objectively due to existing inequalities. For example, degrees obtained from certain universities may be valued more than others even if objectively there seems to be no difference. In addition, even if they are assessed objectively, differences could originate from existing inequalities: a person may have less job experience because of discriminatory employment practices, just like a less educated person may have less education not for the lack of talent and motivation but rather because they belong to a certain population group whose access to education is limited. The last example leads to another very important dilemma: in our application example, where do we draw the line when assessing fairness? If two groups have very different levels of education due to systematic discrimination and yet a certain education level is essential for assessing the candidate's suitability for a certain job, is the algorithm that is based on that desired education level and will consequently select more people from the group with higher average higher education level (the „privileged" group) unfair? While from the societal point of view this would mostly be considered unfair, it is very difficult to decide whether the party that uses the algorithm is responsible for ensuring the fairness that goes beyond the scope of its hiring algorithm (equal chances for that particular job based on the education level or a wider scope of equal educational opportunities) and whether it is responsible for "remedying" the educational inequalities artificially by favoring individuals from the underprivileged group (in the case of gender, the EU law would explicitly permit this (Art. 21, p. 2 of the Fundamental Rights Charter[8])).

The most important thing to realize is that individual fairness and group fairness are generally contradictory. Despite the ongoing philosophical discussion on whether those two concepts are in fact separate concepts or mainly a matter of worldview[9], in practical settings they are clearly distinct.

MEASURING FAIRNESS

In response to the rising awareness of AI-related fairness issues, the research community has developed a wide range of quantitative fairness and bias metrics[10] many of which are already integrated open-source toolkits[11,12,13,14], which assess group fairness, individual fairness or a balanced combination of the two. However, as evidenced by a large number of proposed and applied fairness metrics, there exists no single "ideal" fairness measure due to the challenge of defining fairness described above.

Equality-in-Outcome Approach

Statistical parity difference (spd) is a group-fairness measure defined as:

$$spd = \frac{n(privileged\ classified\ as\ 1)}{n(privileged)} - \frac{n(underprivileged\ classified\ as\ 1)}{n(underprivileged)}$$

which can also be written as probabilities:

$$spd = P(class = 1|privileged) - P(class = 1|underprivileged)$$

How exactly the distinction between privileged and underprivileged groups is made depends on the context of the application. spd compares the fractions of each group classified as 1 (a positive classification, meaning an invitation to a job interview in our hiring algorithm example). This comparison is independent of the group size. In this metric, perfect equality is reached at a value of zero, which means that the same fraction of people is chosen from both groups. For example, if 20% of the overall population are invited for a job interview, in order to be fair 20% of each group should be invited regardless of the group size. This is often described as a "we-are-all-equal" approach since it assumes that there is no difference between the groups even if there is a difference in the data. One main drawback of spd is that it does not address true or false classification of individuals.

Equality-in-Quality Approach

The average odds difference (aod) is another group fairness measure. In contrast to spd, it does not consider classifications as such (e.g., acceptance rates in our job application algorithm example) but rather the quality of the classifications and measures whether the quality rate defined via the false positive (the fraction that is incorrectly classified as 1) and the positive rate (the fraction that is correctly classified as 1) differs between the groups.

$$aod = \frac{1}{2}[(FPR_{underprivileged} - FPR_{privileged}) + (TPR_{underprivileged} - TPR_{privileged})]$$

Perfect equality is indicated by a zero. This is often described as a "what-you-see-is-what-you-get" worldview since it does not assume that an equal fraction from all groups must be accepted, only that the quality of the predictions should be the same for each group. Under this approach, the algorithm can choose only a very small fraction of one of the groups as long as it corresponds to the underlying data. In other words, this approach assumes that the data itself is not biased and accurately reflects reality, ensuring that no bias is introduced by the AI algorithm.

Besides these two examples, there are many other fairness notions, such as conditional statistical parity, equal opportunity, fairness through unwareness, fairness through awareness and counterfactual fairness (https://arxiv.org/pdf/1908.09635v1.pdf).

Below is a summary of issues related to fairness of AI applications:

1. There is no single concept of fairness.
2. Definitions of fairness often contradict each other.
3. Setting the boundary for fairness issues in an application is non-trivial and ambiguous.
4. All of the above points are tightly coupled with ethical considerations and worldviews that depend on many factors, such as region and culture, and can change over time.

In the remaining part of this paper, we discuss how to overcome those obstacles when certifying AI applications in term of fairness.

CERTIFYING FAIRNESS

When developing a certification process, it is customary to establish general requirements for a broad range of applications and subsequently test if each application meets them. However, due to the above-mentioned intricacies surrounding fairness, this method cannot be applied to certifying the fairness of AI applications. In this case, a much more flexible approach is required. Developers of an AI application cannot simply test their application against the fairness criteria since they first have to define them for their application[15]. This has to be accomplished in a transparent and comprehensible way. Moreover, the details of the application and, most importantly, the setting in which the application is intended to be used must be considered. Questions that need to be answered as a prerequisite for certifying fairness are:

- What are the potential fairness issues in the application context?
- What are the vulnerable (potentially unfairly treated) groups?
- Could there be fairness issues that indirectly (rather than directly) relate to the application?
- Could there be undesirable effects in the long term that are not obvious in the short term?
- In which region is the application intended to be used?
- What is the legal context?
- How fair is the status quo (not using the AI application in question)?
- Which means to prevent fairness issues does the application contain?

LuckyStep/shutterstock.com

This list is only a starting point and in no way exhaustive. Given the wide range of possible issues and the subtlety of this topic, it is impossible to compile a complete list of questions. Rather, the questions have to be adapted to the given context. One of the most important things to realize is that potential problems are not always obvious at first glance. Therefore, even directions that seem to be absolutely prone to fairness issues should be considered.

The matters discussed above are not the concern of the certifying party but rather of the application developer/supplier. From a certifier's point of view, one of the most important concepts is that fairness certification only makes sense if in the end the result communicated to the users is not whether "this application is fair" but rather that "this application is fair in the following sense, etc.," with a clear and easy-to-follow description of how fairness was defined in that specific application and context.

The tasks of the certifying party are:

- Review the fairness approach chosen by the developer. Does it make sense? Have any potentially harmful issues been left out?
- Review the actual measures that were taken (e.g., compute fairness metrics, evaluate processes).
- Review the complete application (potentially including the code if it is available) considering the appropriate definition of fairness.
- Create a final fairness report, which describes the fairness definition(s) used in an easily understandable way and how these definitions were met in the application

The process is depicted in figure 1:

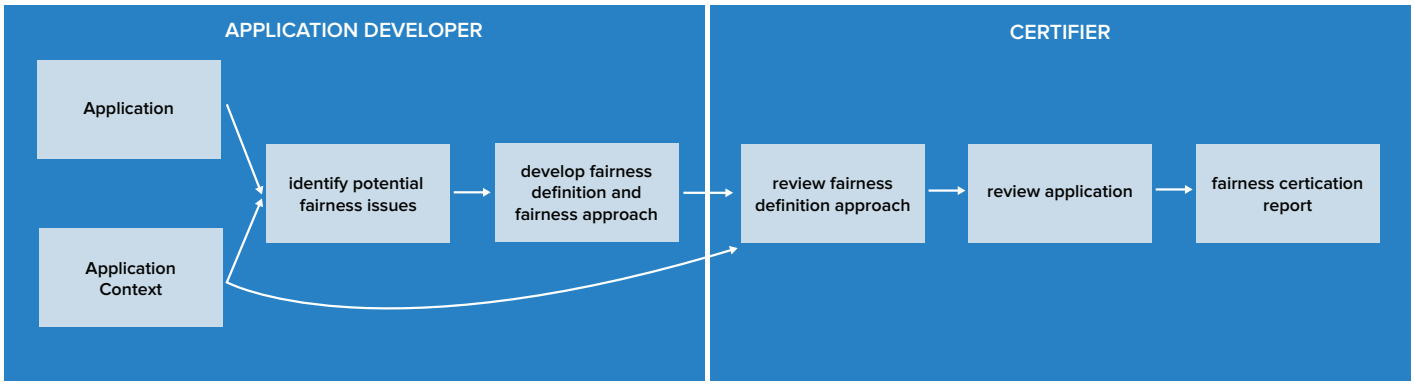

Figure 1: Overview of AI fairness certification from the perspectives of both AI developers and certifiers.

MISSION POSSIBLE!

In this whitepaper, we discussed challenges of certifying fairness aspects of AI applications. We outlined ideas of how these challenges can be addressed and how the fairness certification of AI can be performed.

However, many issues still need to be considered in order to make a routine certification of AI fairness possible, including:

1. Development of reliable (semi-)automatic tools, e.g., tools that automatically compute and visualize fairness and bias metrics.
2. Development of clear guidelines for certifying fairness of AI applications, including international standards.
3. Define case studies that can serve as a reference for certification (e.g.,[16]).

Furthermore, from the legal perspective, there is the problem that the existing quantitative approaches (such as the fairness metrics) are not always in line with how courts deal with fairness and discrimination issues ("A clear gap exists between statistical measures of fairness and the context-sensitive, often intuitive and ambiguous discrimination metrics and evidential requirements used by the Court [European Court of Justice]." [17]). This issue needs to be addressed in order to achieve legal certainty for AI producers.

Being able to independently test and certify AI-applications is crucial not only to avoid legal issues, but also to generate trust in the applications, which in turn is necessary for widespread acceptance of AI.

Since many organizations worldwide are currently working on these topics, we are confident that soon it will be possible to independently certify the fairness of AI applications in a transparent and objective way. Nevertheless, the continuous development of new AI techniques will require constant adaptation of certification procedures.

sdecoret/shutterstock.com

LITERATURE

1 https://www.iais.fraunhofer.de/content/dam/iais/fb/Kuenstliche_intelligenz/ki-pruefkatalog/202107_KI-Pruefkatalog.pdf (in german)

2 Winter, Philip Matthias, et al. “Trusted Artificial Intelligence: Towards Certification of Machine Learning Applications.” arXiv preprint arXiv:2103.16910 (2021).

3 https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52021PC0206

4 https://fra.europa.eu/en/publication/2019/data-quality-and-artificial-intelligence-mitigating-bias-and-error-protect

5 https://fra.europa.eu/en/publication/2020/artificial-intelligence-and-fundamental-rights

6 #BigData: Discrimination in data-supported decision making https://fra.europa.eu/sites/default/files/fra_uploads/fra-2018-focus-big-data_en.pdf

7 Verma, Sahil, and Julia Rubin. “Fairness definitions explained.” 2018 ieee/acm international workshop on software fairness (fairware). IEEE, 2018

8 https://www.europarl.europa.eu/charter/pdf/text_en.pdf

9 Binns, Reuben. “On the apparent conflict between individual and group fairness.” Proceedings of the 2020 conference on fairness, accountability, and transparency. 2020.

10 S. Verma and J. Rubin, “Fairness Definitions Explained,” 2018 IEEE/ACM International Workshop on Software Fairness (FairWare), 2018, pp. 1-7, doi: 10.23919/FAIRWARE.2018.8452913.

11 Bellamy, Rachel KE, et al. “AI Fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias.” arXiv preprint arXiv:1810.01943 (2018).

12 https://aif360.mybluemix.net/

13 https://www.microsoft.com/en-us/research/uploads/prod/2020/05/Fairlearn_White-Paper-2020-09-22.pdf

14 https://fairlearn.org/

15 https://www.iais.fraunhofer.de/content/dam/iais/fb/Kuenstliche_intelligenz/ki-pruefka-talog/202107_KI- Pruefkatalog.pdf (in german)

16 https://www.technologyreview.com/2021/02/11/1017955/auditors-testing-ai-hiring-algorithms-bias-big-questions-remain/

17 Wachter, Sandra, Brent Mittelstadt, and Chris Russell. “Why fairness cannot be automated: Bridging the gap between EU non-discrimination law and AI.” Computer Law & Security Review 41 (2021): 105567.

PUBLISHER

Know Center GmbH

Research Center for Data-driven Business & Big Data Analytics

Inffeldgasse 13/6, A-8010 Graz

This publication is part of an ongoing series by Know-Center on certification and conformity assessment of AI applications within the “Trust Your AI” initiative.

trustyourai@know-center.at

https://trustyour.ai/en/