Fairness by awareness? On the inclusion of protected features in algorithmic decisions

This is an accepted preprint version!
You find the revised document here: https://www.sciencedirect.com/science/article/abs/pii/S0267364922000061?via%3Dihub.

Hanna Hoffmann[a,*], Verena Vogt[b], Marc P. Hauer[c], Katharina Zweig[d]

[a]*Institute for Information, Telecommunication and Media law (ITM) - WWU Münster, Germany*
[b]*Institute for Information, Telecommunication and Media law (ITM) - WWU Münster, Germany*
[c]*[OrcID 0000-0002-1598-1812], Algorithm Accountability Lab (AAL) - TU Kaiserslautern, Germany*
[d]*[OrcID 0000-0002-4294-9017], Algorithm Accountability Lab (AAL) - TU Kaiserslautern, Germany*

Abstract

AI decisions are increasingly determining our everyday lives. At present, European anti-discrimination law is process-oriented; it prohibits the inclusion of sensitive data that is particularly protected. However, especially in the context of AI decisions, constellations can be identified in which the inclusion of sensitive characteristics will lead to better and sometimes even less discriminatory result. A result-oriented approach, therefore, might be a more fitting strategy for algorithmic decision making.

In this paper we examine the legal framework for including sensitive features in a Support Vector Machine for a fictitious scenario and discuss the resulting challenges in practical application. It turns out that generally ignoring sensitive features - as has been the practice up to now - does not seem to be a fitting strategy for algorithmic decision making. A process-oriented procedure only supposedly comes closer to individual case justice: If one assumes that fewer errors occur when protected characteristics are included, individuals will ultimately also be assessed incorrectly less often, especially when one protected group is more prone to errors than the other.

This paper aims to support the current debate about legal regulation of algorithmic decision making systems by discussing a perspective often neglected.

Keywords: Fairness, Discrimination, Machine Learning, Legal policy, Algorithmic Decision Making

1. Introduction

In the age of Big Data analyses, evaluating large amounts of data is becoming possible at ever lower costs [1]. The successes of so-called ”artificial intelligence”, for example in the area of image recognition or the processing of spoken language, are also increasing the desire to use the seemingly objective calculations for more delicate decisions: For instance, according to a Reuters report, Amazon has attempted to develop automated job applicant assessment software using machine learning. After a few months, however, the system was found to favor male applicants. The project has since been discontinued.[1] Such algorithms receive various information about a person and, according to a set of rules, then calculate a probability value, e.g., how likely a person is to leave the company again within two years. In some of these tools, the decision rules are set manually, while others ”learn” the rules from past data using a

*Corresponding author
Email addresses: hanna.hoffmann@uni-muenster.de (Hanna Hoffmann), verena.vogt@uni-muenster.de (Verena Vogt), hauer@cs.uni-kl.de (Marc P. Hauer), zweig@cs.uni-kl.de (Katharina Zweig)

[1]Dastin, Amazon scraps secret AI recruiting tool that showed bias against women, Reuters, 11.10.2018, available under: https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G (last accessed on 22.10.2021).

machine learning method: this involves searching a set of data for characteristics of those employees who frequently resign and those who have not resigned or have remained with the company for as long as possible. For example, it is known – and this can also be extracted from the data as a correlation by machine learning methods – that young women are more likely to quit a job than older men.[2] The correlations observed ("learned") in this way are then stored as a set of decision rules and used to assess the risk of further employees. It can be observed that sensitive characteristics are taken into account in the decision-making process. It was previously assumed that not using or not knowing such information would prevent a decision maker from acting in a discriminatory manner.[3] Accordingly, this information should generally be ignored, but even with such a requirement discrimination can occur in other ways. According to the Reuter's report mentioned above, this was also the case with Amazon. Although the software did not know whether a woman's or a man's resume was available for evaluation, previous hiring practices indicated that women were less likely to be hired. This tendency was picked up by the machine. For example, people who had graduated from a women's college or who said they had run a club for women only were rated lower. After Amazon became aware of this, the software was discontinued entirely; until then, it had not been used on a large scale, the report said.

Up until now, it has therefore been common practice in the context of selection decisions not to take protected characteristics into account – such as those of Article 3 (3) Sentence 1 of the German Constitution (Ger. Const.), Article 21 (1) of the EU Charter of Fundamental Rights (CFR), and, at the level of German sub-constitutional law, Section 1 of the German General Equal Treatment Act (GET) and Article 9 (1) of the General Data Protection Regulation (GDPR) – as a matter of principle, in order to ensure that the regulations in question are not violated. A selection or decision-making process is perceived as fair if it proceeds independently of these protected characteristics.

However, the use of AI now raises the question of whether our previously almost exclusively process-oriented understanding of fairness and justice, should gradually evolve into a result-oriented view. By the term "process-oriented" we mean those aspects that relate directly to the decision-making process, i.e., how the decision is reached. The term "result-oriented" refers to those aspects that are linked to the result and do not take into account how the decision was made [1]. Based on this, we will examine whether a protected characteristic may be systematically included in a decision-making process for a "good cause", or should be included if the result achieved this way would reduce errors at the outcome level, which would be particularly detrimental to persons with precisely these protected characteristics. It should be noted at this point that, in addition to the legal issues, details of the specific situation make an assessment of this question complex in reality.

2. Technical background to machine learning

As briefly outlined in the introduction, machine learning methods can derive decision rules from a data set. We demonstrate the mechanism using a fictitious example: for the evaluation of job applications, an AI is to be created that allows predicting which applicants are most likely to be suitable for the job based on the information mentioned in the resume. In order to do this, the applications to similar jobs in the same company from the last few years and the information about whether the applicants were hired or not are collected. In the following, we refer to this collection of information as the (training) data set and the information whether a person was hired or not as the so-called ground truth. In the application process, the AI system is to be used to support the HR department in deciding which of the applicants should be invited for a personal interview.

The development of such AI systems is always based on at least one result-oriented quality measure and at least one fairness measure with which the decisions are evaluated.

A system that produces as few wrong decisions as possible can be called good – there are about two dozen different formulas for measuring such "goodness" alone [2]. A system that does not judge the minority group significantly better or worse than the group that represents the majority is called fair. One possible fairness measure to meet this requirement is called Conditional Independence [3]. According to this fairness measure, the system would be fair if the ratio of people invited to an interview out of the applicants considered suitable is the same for both groups. Suppose 1,000 applicants applied for a job, 700 women and 300 men. We now assume that 40% of the men (i.e.

[2]Nordic Council of Ministers, Labour Market Mobility in Nordic Welfare States, p. 207.
[3]BAG, 5.2.2004 - 8 AZR 112/03, NZA 2004, 540 (544).

120) and 60% of the women (i.e. 420) meet the job suitability requirements. Now, if 21 of the eligible 420 women (5%) are recommended by the system for invitation, then 6 of the sufficiently qualified 120 men (5%) must also be recommended by the system for it to be considered fair in terms of Conditional Independence.

One of the methods that can be used to derive decision rules from such data sets is called a Support Vector Machine (SVM): A SVM is a mathematical learning procedure that gets a data set that (in simple cases) splits into two groups; in the criminal justice context, for example, recidivist and non-recidivist offenders. The software then looks for a dividing line that separates the two groups as well as possible, i.e., the points of the two groups should preferably be on only one of the two sides and as far away from the dividing line as possible. Technically, the ”dividing line” is only a line if there are only two other pieces of information about the individuals beyond recidivism (e.g., age and gender). If there are three pieces of information, the dividing line becomes what is called a hyperplane, which can be thought of as cutting through a cube. A ”hyperplane” in general determines where exactly the boundary between the two groups is, for all possible combinations of information.

Figure 1 shows a visualization of a data set. The data points could be, for example, the successfully hired applicants (green, square) and the not successfully hired (red, triangular). The properties by which the data points are plotted are ”years of expertise in the job” and ”number of months without employment since completion of advanced training.” Now, a dividing line is to be introduced. For new data points, it will help to decide to which group the applicant belongs.

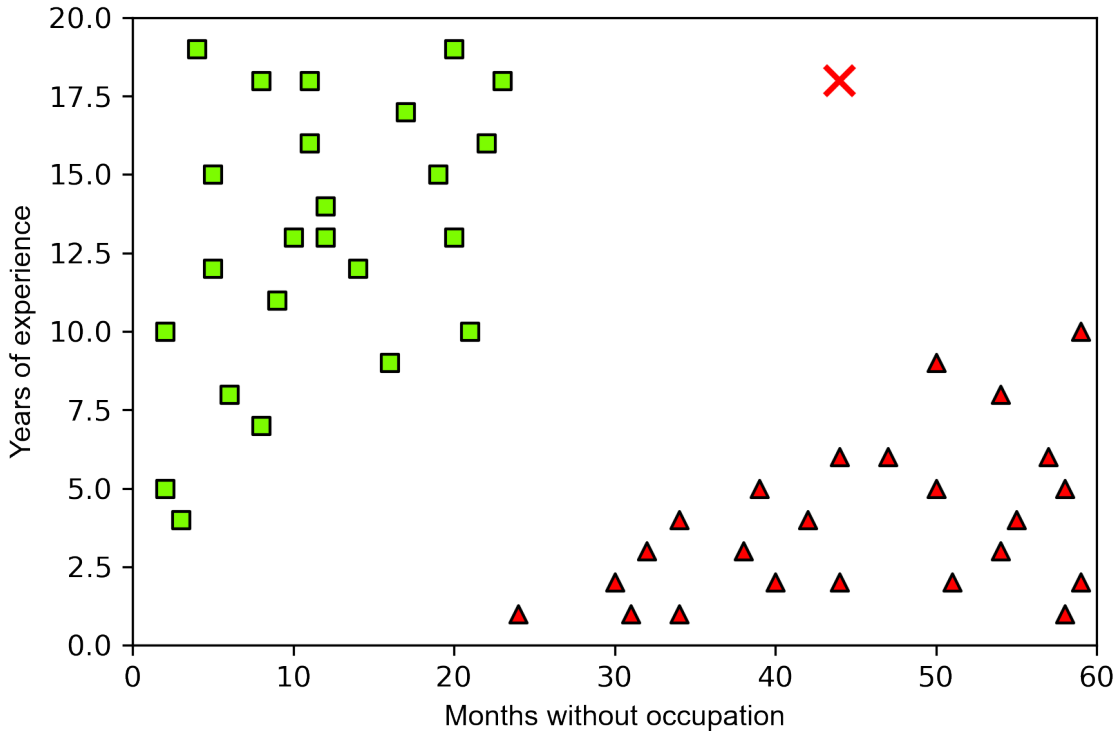

Figure 1: It is unclear whether the data point marked here with an X is a person who is likely to be successfully recruited or not. An SVM is intended to provide an assessment.

In this example, a number of dividing lines can be drawn by hand without difficulty, but this also means that there are a number of possible dividing lines that appear equally good at first glance (Figure 2, left). Different separators may group a new data point in different ways (Figure 2, right).

A Support Vector Machine can solve the problem by mathematically calculating an optimal separation line be-

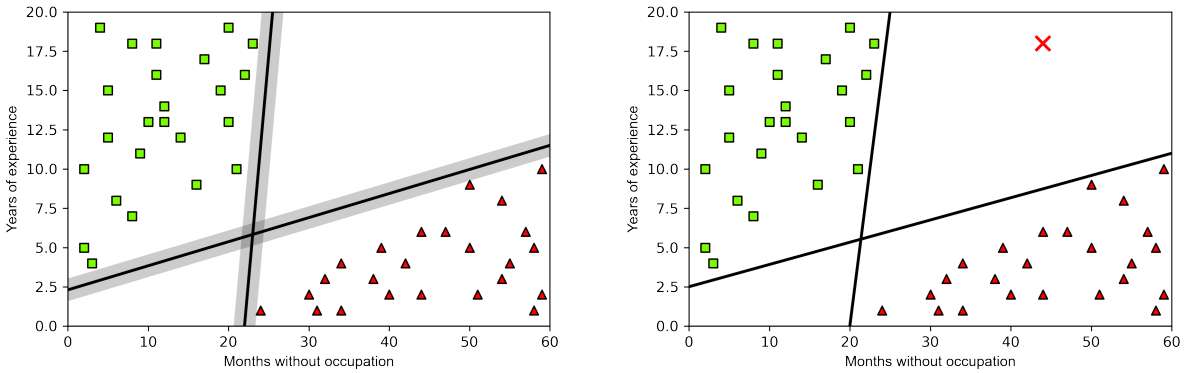

Figure 2: Depending on the choice of the dividing line, the data point marked with X would be judged as a successful or unsuccessful applicant.

tween the groups that maximizes the distance to adjacent points. The points that are decisive for this distance are the support vectors that give the method its name (Figure 3).

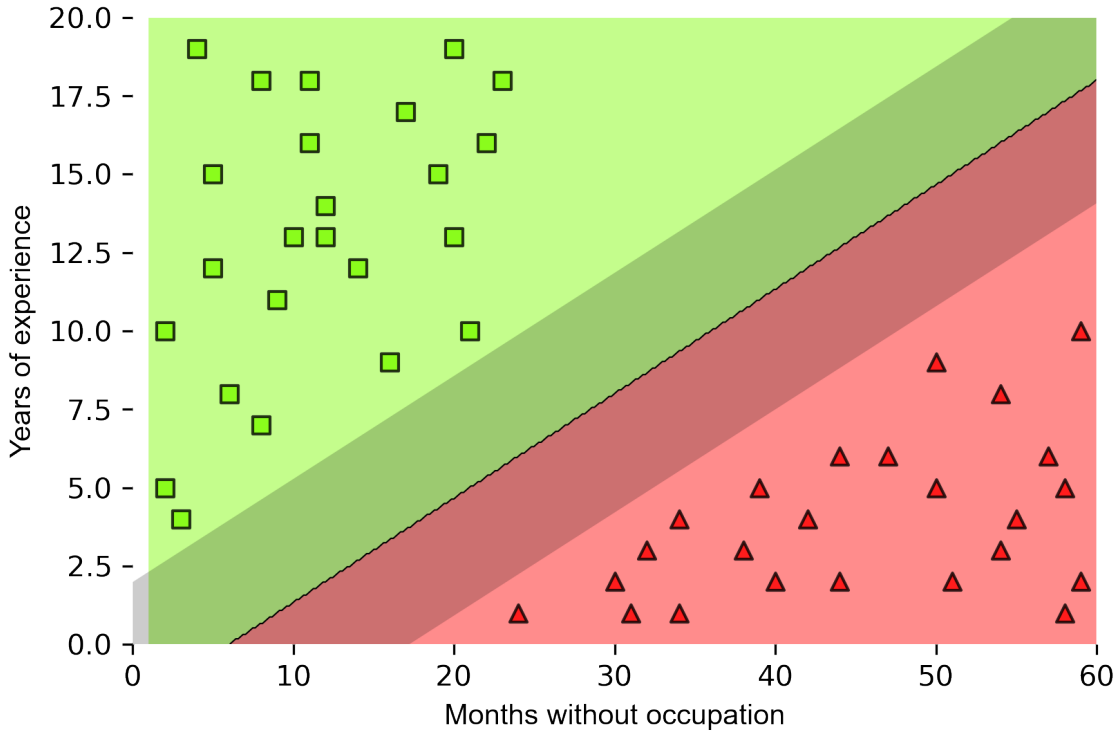

Figure 3: An optimal dividing line maximizes the distance to the nearest points. This is visualized here by the distance to the dividing line.

Once this dividing line (or hyperplane in the more general case) has been calculated, it is used as a decision rule for the further process: New points are evaluated according to whether they fall on one side of the boundary or the other, and this determines the output of the estimation.

Even though we are specifically looking at the SVM method here, it is important to emphasize that all machine

learning methods basically follow the same procedure: They analyze a data set, looking at the information available for each individual and how they have been estimated in the past. Building on the patterns and correlations within the data set, they create a decision rule that dictates how individuals should be assessed in the future based on their data. Like in our example, this could be done by categorizing them into one of two classes.

3. Use Case

In reality, the groups of "successful" and "unsuccessful applications" are rarely so easy to separate: There is often conflicting information. Figure 4 shows a data set that is more realistic; the two groups overlap. Here we use the concept of SVM explained above, extended to allow for misjudgments. It now attempts to minimize the number and magnitude of errors whenever possible.

The training result is a dividing line which separates successful and unsuccessful applications as best as possible and from which a mathematically optimal decision rule can thus be derived (Figure 4).

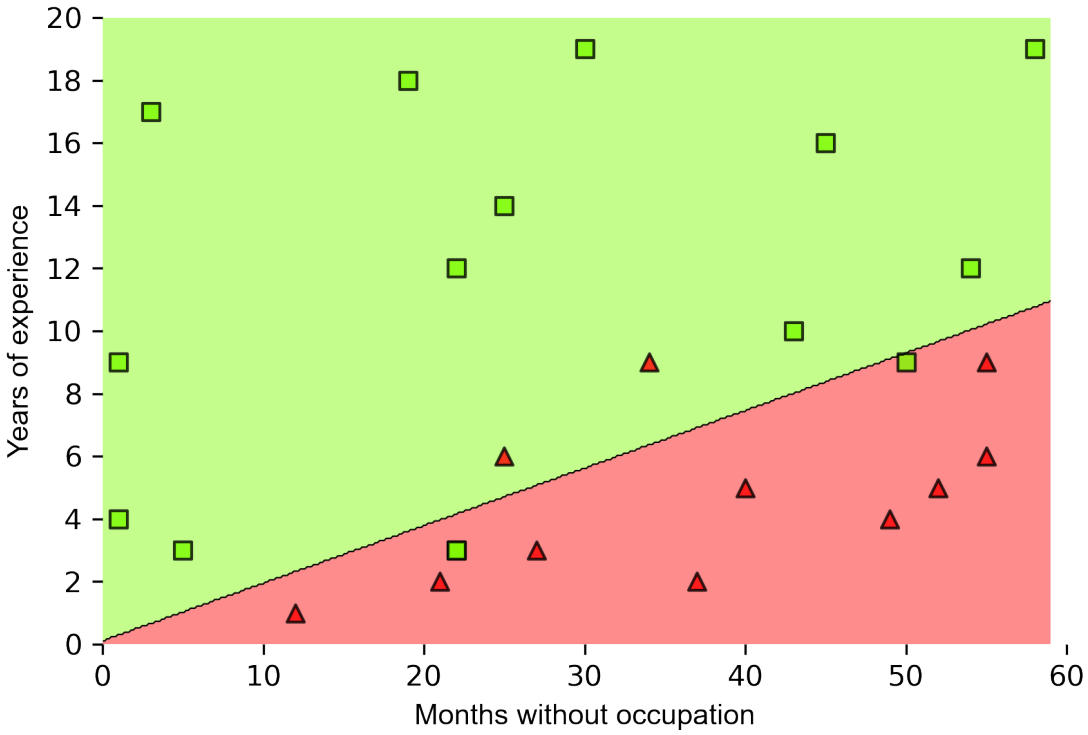

Figure 4: Fictive training data set of people who have been hired, respectively not hired, together with the boundary between them calculated by an SVM, which has the largest possible total distance to the surrounding points.

Four mistakes are made on the training data: two good applicants are wrongly assigned to the group of bad applicants by the SVM, and two bad applicants are wrongly invited.

This is the optimal solution under the condition that the protected property (gender in this case) cannot be used, i.e., the optimal solution under the process-oriented view (blind model). Under the result-oriented view, which considers the protected property, the situation is different: Figure 5 shows the same data set, but now additionally marks the female applicants with a circle. The figure shows that, although the procedure was agnostic about the gender of applicants, a gendered pattern is nevertheless evident: The applicants incorrectly classified as presumptively unsuccessful are all women, and those incorrectly classified as presumptively successful are all men.

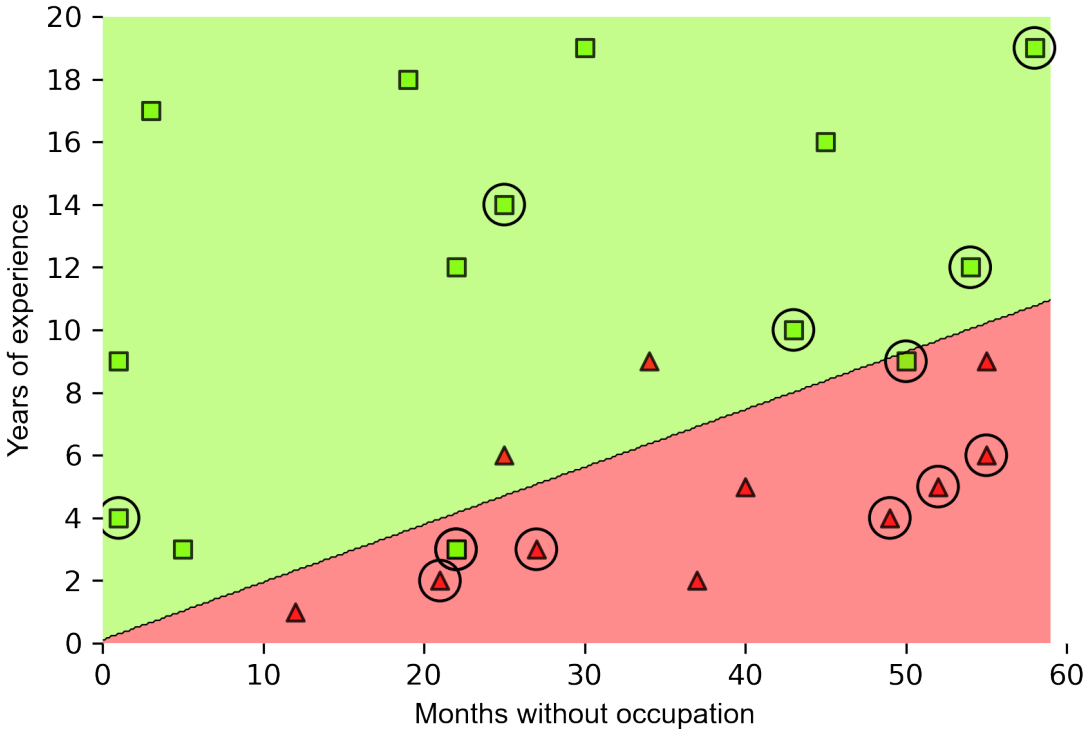

Figure 5: Data points representing women are circled.

In fact, this can always happen when protected properties are withheld from an algorithm which also have an influence on behavior or treatment [4]. As can be seen on the Amazon example, this is not only a theoretical problem. It has also been demonstrated for the COMPAS system, which is still used in American courts today to calculate how likely it is for criminals to recidivate at different stages of a trial.[4]

In contrast, this worst-case scenario can be avoided in our fictitious example if the training data set is split by gender and two independent SVMs are created (**differentiating model**). In this case, a separate, mathematically optimal decision rule is derived for each subgroup (Figure 6), whereby, in this example, no errors are generated. The fictitious example shows that the consideration of protected features can provide clearly better outcomes from the result perspective. It is important to emphasize that the machine learning method used for this does not affect the demonstration of this effect. Similar examples can be constructed for other methods.

The two incorrectly categorized women are not invited in the gender-blind model only because of that very blindness, whereas they are identified as ”suitable candidates” according to the differentiating SVM. At the same time, it can be observed that the dividing line between the decision to invite a candidate is lower for women when considered separately compared to when considered jointly, while it is higher for men. The differentiated analysis increases the requirements that men must meet in order to be considered suitable. It follows that there is now a group of male applicants who would meet the requirements under the blind model, but who do not meet the requirements when the differentiating model is used.

In what follows, we base our legal analysis on the first scenario, but would like to reiterate at this point that the disadvantaged gender depends on the specific situation and not generally on whether or not gender is taken into

[4] Angwin/Larson/Mattu/Kirchner, Machine Bias: There’s Software Used Across the Country to Predict Future Criminals. And It’s Biased Against Blacks., ProPublica, 23.03.2016, available under: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing (last accessed on 22.10.2021).

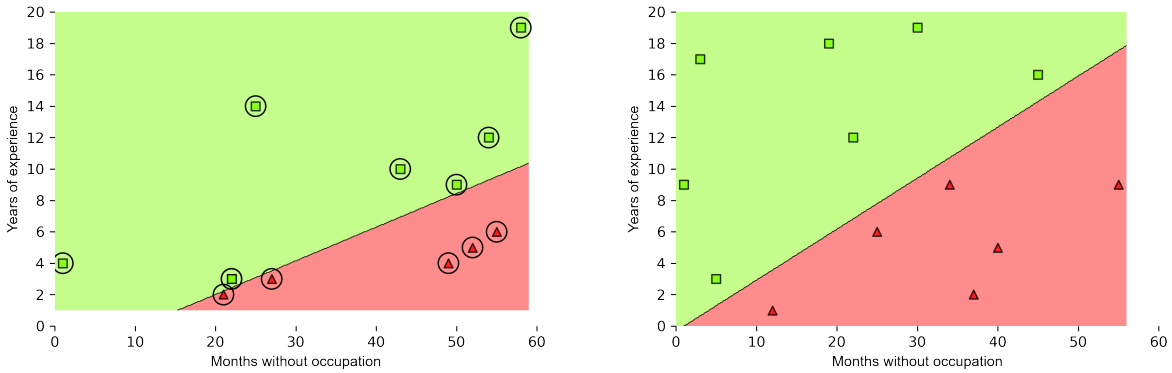

Figure 6: Separate SVMs for women (left) and men (right).

account to generate one (or more) SVMs.

4. Legal background

For the legal analysis, this scenario has to to be examined with regard to data protection law and anti-discrimination law.

4.1. Data Protection Law

The calculations of the SVMs are based on the evaluation of personal data. The processing of personal data is not only protected by Art. 8 CFR, but also by the GDPR.

4.1.1. Art. 8 CFR

According to Art. 51 CFR, the provisions of the Charter are addressed to the institutions and bodies of the Union and to the Member States only when they are implementing Union law. Nevertheless, the Charter has also a significant effect among private individuals: On the one hand there are indirect effects such as interpretation in conformity with fundamental rights and the non-applicability of norms of private law. The CJEU interprets the relevant provisions in private law cases in the light of the fundamental rights.[5] On the other hand, certain fundamental rights are directly binding on private individuals. However, this does not apply to Art. 8 CFR [5, Breuer, § 25 para. 46] [6, Streinz, Art. 8 CFR para. 6] [7, Jarass, Art. 8 para. 3]. According to Art. 8 (1) CFR, everyone has the right of protection of personal data concerning him or her. Under Art. 8 (2) Sentence 1 CFR, personal data must be processed fairly for specific purposes and on the basis of the consent of the person concerned or some other legitimate basis laid down by law. Moreover, according to Art. 8 (2) Sentence 2 CFR, everyone has the right of access to data which has been collected concerning him or her and the right to have it rectified. These requirements must be observed in particular when interpreting the GDPR.

4.1.2. GDPR

The GDPR is linked to the CFR in many places and is to be interpreted in the light of the Charter. Although the protective purpose of the GDPR is not primarily oriented towards the protection against discrimination, the protection of personal data can have an indirect impact on discrimination. The GDPR differentiates between ”simple” personal data and special categories of personal data. According to Art. 4 no. 1 GDPR, personal data is any information relating to an identified or identifiable natural person; whereby an identifiable natural person is one who can be

[5] See CJEU, C-360/10 – Sabam, 26.2.2012 para. 52; C-426/11 – Alemo-Herron, 18.7.2013 para. 30; C-131/12 – Google Spain, 13.5.2014, para. 68 f, 74; C-580/13 – Coty, 16.7.2015 para. 34.

identified, directly or indirectly, in particular by reference to an identifier such as a name, an identification number, location data, an online identifier or to one or more factors specific to the physical, physiological, genetic, mental, economic, cultural or social identity of that natural person. Special categories of personal data are, according to Art. 9 GDPR, data revealing racial or ethnic origin, political opinions, religious or philosophical beliefs, or trade union memberships, and the processing of genetic data, biometric data for the purpose of uniquely identifying a natural person, data concerning health or data concerning a natural person's sex life or sexual orientation. Since processing sensitive data needs special protection, the processing of sensitive data is prohibited pursuant to Art. 9 (1) GDPR unless an exemption provision of Art. 9 (2) GDPR applies. Furthermore, a justification according to Art. 6 (1) GDPR is required. The precise legal requirements depend on the specific individual case. Depending on what kind of personal data are processed and depending on how the circumstances are structured, the lawfulness of the processing will be determined. With regard to the scenario, it can be noted that the gender feature does not belong to the special category of personal data, so the lawfulness is only based on Art. 6 GDPR. According to Art. 6 (1) GDPR, data processing is lawful if one of the justification grounds listed in para. 1 applies. Even if these requirements are significantly less strict than those of Art. 9 GDPR, they can be a hurdle for the processor. As the GDPR only applies to personal and pseudonymized data, but according to Recital 26 not to anonymous information, some therefore suggest anonymizing the personal data before processing them[8, pp.183-201, pp.198]. Anonymization means modifying personal data in such a way that the individual information about personal or factual circumstances can no longer be assigned to an identified or identifiable natural person, or can only be assigned to an identified or identifiable natural person with a disproportionate amount of time, cost and effort. Therefore, anonymization approaches are difficult to implement, as there are efficient ways to de-anonymize single data points for which at least a little information is known [9]. Additionally, even effective anonymization might not be sufficient for protecting an individual from personalized attacks. Even if it is not possible to identify an individual in a data set, in many cases it is possible to assign him or her to a small group with very specific characteristics which can be targeted instead of a precise individual [10]. Depending on the specific case, different grounds for justification must be used (and, if necessary, grounds for exemption according to Art. 9 (2) GDPR, if special categories of data are processed). In our scenario, consent of the data subject (Art. 6 (1) lit. a GDPR), the fulfilment of a contractual obligation (Art. 6 (1) lit. b GDPR) or a legitimate interest pursued by the controller or by a third party (Art. 6 (1) lit. f GDPR) could be applied depending on the circumstances of the case.

4.2. Anti-Discrimination Law

Legal provisions regarding anti-discrimination can be found in the EU-primary legislation (Art. 21 and 23 CFR) as well as in secondary legislation (anti-discrimination directives).

4.2.1. Art. 21 and 23 CFR

The European Charter of Fundamental Rights (CFR) contains two articles on non-discrimination. The first is Art. 21 CFR, which supplements the general principle of equality (Art. 20 CFR) with special prohibitions of discrimination based on any ground such as sex, race, colour, ethnic or social origin, genetic features, language, religion or belief, political or any other opinion, membership of a national minority, property, birth, disability, age or sexual orientation. These groups of discrimination are based on provisions of international law (e.g. Art. 14 of the European Convention on Human Rights) and on a common constitutional tradition of the European member states [11, Hölscheidt, Art. 21 para. 2.]. Art. 21 CFR standardizes in Art. 21 (1) CFR a comprehensive general prohibition of discrimination as well as a prohibition of discrimination on grounds of nationality in Art. 21 (2) CFR. The second is Art. 23 CFR, which ensures equality between men and women. Since the provisions of Art. 21 (2) and Art. 23 CFR are more specific than those of Art. 21 (1) CFR, it is secondary to both of the previously mentioned regulations [11, Hölscheidt, Art. 21 para. 31.]. As a prohibition, Art. 21 CFR does not contain any duties of protection, but Art. 19 (1) of the Treaty on the Functioning of the European Union (TFEU) empowers the EU legislature to take "appropriate" precautions against discrimination (e.g. directions) [11, Hölscheidt, Art. 21 para. 56.]. In contrast, the claim of gender equality enshrined in Art. 23 CFR is linked to a protection mandatehttps://www.beck-shop.de/pechstein-nowak-haede-frankfurter-kommentar-euv-grc-aeuv/product/18037219. But since there can be no actual equality between people in general and between members of different genders in particular, the claim of Art. 23 CFR can basically only be a matter of avoiding unjustified discriminatory differentiation on the basis of gender [12, Art. 23 GRC Rn. 23.]. Therefore,

the addressees must refrain from any discrimination and are responsible to create a situation free of discrimination [11, Art. 21 para. 56.]. In this context, discrimination is understood to mean disadvantageous unequal treatment, whereby not every unequal treatment already constitutes discrimination within the meaning of Art. 23 CFR [13, Kugelmann, § 160 para. 42]. A distinction is made between direct discrimination, which is directly linked to the prohibited attribute, and indirect discrimination [13, Kugelmann, § 160 para. 44]. The latter is not linked to a prohibited attribute and thus gives the appearance of neutrality, but it leads to a certain group being disadvantaged [13, Kugelmann, § 160 para. 44]. However, the rights under Art. 21 and Art. 23 CFR are not mandatory, but may be restricted under certain conditions. In principle, discrimination can be justified in the same way as unequal treatment under Art. 20 CFR [14, Rossi, Art. 21 para.9] [7, Jarass, Art. 21 para. 26]. Thus, unequal treatment is compatible with Art. 21 and Art. 23 CFR if it is objectively justified, based on an objective and reasonable criterion and if the difference in treatment is proportionate to the objective pursued.[6] Furthermore, unequal treatment of women and men may be justified under the case law of the CJEU.[7] Directly, fundamental rights such as Art. 21 and Art. 23 CFR oblige only the EU and pursuant to Art. 51 (1) Sentence 1 CFR they oblige only the member states insofar as they implement EU law. But since law must be interpreted in the light of fundamental rights, national law and the legal relationship between private parties are also affected by this horizontal effect.[8] The CJEU has affirmed the direct binding of private parties for Art. 21 (1) CFR, for example. In principle, however, no concretisation of the norm may be necessary[9] and indications for a direct application must be present [15, p.402]. Furthermore, it is discussed whether the applicability only exists in the case of the obligation of superior private persons in asymmetrical relationships [7, Jarass, Art. 51 para. 42].

4.2.2. Anti-Discrimination Directives

By specifying the primary law in the form of the CFR, the secondary law lays down a variety of concrete prohibitions of discrimination in the form of directives. A fundamental rule of EU law is equality of treatment, which must be observed in every field covered by the European Treaties.[10] Prohibitions on discrimination were harmonised in the EU in Directives 2000/43/EC, 2000/78/EC, 2004/113/EC and 2006/54/EC. They are based on Art. 19 TFEU and implement the prohibition of discrimination from Art. 21 and 23 CFR, which must be enforced by the Member States. [16, Schubert, Art. 28 CFR para. 100] The directives demand that the Member States enact laws to counter disadvantages arising on the grounds of race, ethnic origin, religion, age, disablement, gender or sexual orientation. In Germany, for example, the anti-discrimination directives have been implemented in the General Equality Act (GET), and in Austria in the Equal Treatment Act. Particularly relevant for the mentioned scenario is Directive 2006/54/EC as it addresses the principle of equal opportunities and equal treatment of men and women in matters of employment and occupation.

Directive 2006/54/EC states in Article 2 (1) lit. a with regard to direct discrimination that one person may not be ”treated” worse than another on the basis of his or her sex. Art. 14 (1) lit. a of Directive 2006/54/EC also specifically provides that discrimination is prohibited with respect to ”conditions of access to employment, self-employment or occupation, including selection criteria and recruitment conditions”. With regard to indirect discrimination (Article 2 (1) lit. b), it states that a person must not be subjected to a disadvantage on the basis of his or her sex in comparison with another person. However, it does not define when a disadvantage exists. The same applies to Directive 2004/113/EC, as well as to Directives 2000/78/EC and 2000/43/EC with regard to other protected criteria. Since in the case of direct discrimination, it is not the result but the treatment and thus the procedure that is referred to. The wording thus arguably speaks for a process-oriented understanding. It is therefore a matter of justice in individual cases and the comprehensibility of the decision-making process that matters. Furthermore, from a systematic point of view, the wording must not be understood too narrowly since at least indirect discrimination can still be justified.

[6] cf. CJEU, C-101/12 - Schaible, para. 77; CJEU, C-390/15- RPO, para. 53; [7, Jarass, Art. 20 para. 15]

[7] CJEU, C-236/09 - Test-Achats, para. 28; [7, Jarass, Art. 23 para. 6]

[8] CJEU, C-423/04 - Richards, Slg. 2006, I-3585 para. 23 f.; [7, Jarass, Art. 23 para. 5]

[9] CJEU, C-569/16 – Bauer, 6.11.2018, ECLI:EU:C:2018:871, para. 84; CJEU, C-684/16 – Max-Planck-Gesellschaft, 6.11.2018, ECLI:EU:C:2018:874, para. 74.

[10] CJEU, C-122/17 – David Smith, 7.8.2018 para. 47; Fuchs/Cornelissen, EU Social Security Law, Regulation (EC) No 883/2004, Article 4 Equality of treatment, para 13.

5. Does the principle ”fairness by awareness” constitute discrimination?

From a purely factual point of view, male applicants in the said scenario have to meet higher requirements than female applicants. A male applicant to whom a higher cut-off line applies is therefore disadvantaged compared to female applicants under a process-oriented understanding. The fact that the minimum requirements are determined solely on the basis of gender suggests a direct discrimination within the meaning of Article 2 (1) lit. a. However, since the minimum requirements are determined on the basis of the information in the respective data set and other information could lead to the requirements for female applicants being higher than those for male applicants, an indirect discrimination within the meaning of Article 2 (1) lit. b could also be assumed. Indirect discrimination is also indicated by the fact that men are not expected to meet higher standards because they are men, but rather because their data set dictates this. One could therefore argue that the distribution of characteristics, and not the sex characteristic alone, dictates the minimum requirements. Therefore, the scenario constitutes indirect discrimination (Article 2 (1) lit. b). Thus, it is crucial whether the use of the different decision formulas is objectively justified.

6. Justification of the model

In the following, we will examine whether the use of the differentiating model can be justified and how this plays out in reality.

6.1. Comparison with the case law of the CJEU

An obvious first step is to compare the present scenario with case law on insurance rates. Until the CJEU ruling in March 2011[11], it was common practice in some areas for women and men to pay different rates for certain insurance policies. This was possible because among others, the German legislature exploited a leeway provided by the so-called Gender Directive[12] and, with § 20 (2) Sentence 1 GET, relativized the regulation of gender-neutral rates enshrined in § 19 (1) No. 2 GET [17]. This was justified with different high risks, which was proven statistically.[13] For example, women are less frequently involved in traffic accidents than men [18, pp.348], while men have a shorter life expectancy and therefore receive lower pension payments [18, pp.348]. The CJEU ruled that a rule allowing Member States to maintain different levels of premiums and benefits indefinitely was not compatible with Articles 21 and 23 CFR.[14] This decision of the CJEU could also provide direction for the present scenario. After all, the initial situation of the insurance case was also that men and women were evaluated separately and had to pay different premiums based on the results of this evaluation in order to achieve potentially ”fairer” results. The freedom of contract and entrepreneurial freedom, Art. 16 CFR, had to be weighed against the protection against discrimination of Art. 21, 23 CFR [19, p.297]. In European law, freedom of contract is not explicitly standardized, but is indirectly derived from the overall view of Union law, the case law of the CJEU, and the statements of the Commission [20, p.38]. In the area of insurance, however, the principle of solidarity was added, which is not of decisive importance in our concrete scenario. If one considers solidarity as a principle for society as a whole, in which it is generally a matter of people with different risks standing up for one another, then differentiation is precisely not wanted [21]. However, the CJEU's ruling was not received entirely uncritically in publications (see for example [22, p.423] [23, p.16]). On the one hand, it was criticized that a gender-specific approach was not sufficiently meaningful for a risk assessment; rather, it would depend on the individual ways of life and behavior.[15] Although the approach of individual consideration was approved, it was noted that there are certain risks that can only be assigned inherently to one sex, such as diseases from which only one sex can suffer (for example, cervical or prostate cancer) and must therefore only be included in the risk assessment for the respective sex [24, p.428] [25, p.170] [26, p.701]. The decision therefore fails to recognize the crucial difference between sex differentiation and sex discrimination [22, p.431] [26, p.701] [25, p.170].

[11] CJEU, C-236/09, 1.5.2011, ECLI:EU:C:2011:100.

[12] Directive 2006/54/EC.

[13] CJEU, C-236/09, 1.5.2011, ECLI:EU:C:2011:100, para. 27.

[14] CJEU, C-236/09, 1.5.2011, ECLI:EU:C:2011:100, para. 32.

[15] The CJEU did not express this directly, but certainly the opinion of Advocate General Juliane Kokott of 30.9.2010 on Case C-236/09, paras. 60-63; Armbrüster VersR 2010, 1571 (1581 f.).

The EU Commission also recognized this problem and took the position that, with regard to diseases that exclusively or mainly affect men or women, it continues to be possible for insurers to offer gender-specific insurance products or gender-specific options within insurance contracts under the conditions of Art. 4 (5) of Directive 2004/113/EC.[16] Moreover, the use of gender as a factor for risk classification is not generally prohibited if it is used in the calculation at the aggregate level and as long as it does not lead to a distinction at the individual level.[17] There are risk factors, such as health status or family history, on the basis of which differentiation is possible and insurers must take gender status into account when assessing them, given certain physiological differences between men and women.[18] A family history of breast cancer, for example, does not affect a man's health risk to the same extent as a woman's, so it must be known whether the person is a woman or a man to evaluate this effect.[19] Furthermore, for example obesity is a risk factor whose measure is the waist-to-hip ratio, which is different for men and women.[20] Secondly, the court criticized the minimalist justification effort, which completely lacks a proportionality test and thus gives the impression that the contractual freedom of the insurers is not taken into account [27, p.574].This is all the more serious since, according to Art. 119 (1) TFEU, the economic policy of the Member States is based on the principle of an open market economy with free competition, which necessarily presupposes the private autonomous decision of citizens to bind themselves by contract [25, p.166]. In this respect, it seems questionable to what extent equal treatment bears the concept of justice if this leads to the result that one group must share the risks of the other group without even potentially being exposed to them [26, p.701] [25, p.107]. Applied to our scenario, the question therefore arises as to whether it is permissible to assess female applicants according to a procedure if it is established that they are systematically disadvantaged as a result, or whether it is not rather permissible to include protected characteristics in the decision-making process if it is established that this produces a better result overall. The proportionality of a decision-making process that takes gender characteristics into account will be examined in the following.

6.2. Justification of the scenario

For unequal treatment to be justified, first there must be a legitimate reason for differentiation. The measure must then be suitable for achieving this purpose and be necessary. Finally, it is examined whether the measure is appropriate in relation to the disadvantage [28, Baumgärtner, § 3 AGG para. 88].

In our scenario, the legitimate purpose is the employer's interest in selecting the best possible applicants. It is not contrary to the process-oriented approach of Directive 2006/54/EC to use a result-oriented aspect as part of the justification. However, the requirement of the law must be taken into account in the context of the subsequent weighing of interests.

The method is also suitable. Any measure that contributes to the realization of the purpose is suitable [28, Baumgärtner, § 3 AGG para. 97]. By using different decision formulas, a better result is achieved than would be possible with a common calculation method.

Furthermore, the procedure is also necessary because no equally suitable milder method is available. On the contrary, it is clear that more errors would have been made by using a common decision formula. It would be conceivable to have the application documents reviewed by a natural person. However, this approach would not be equally appropriate. On the one hand, from the point of view of the company's effectiveness, the use of employees is significantly more cost-intensive and time-consuming than algorithms, although it should be pointed out at this point that the principle of equal treatment must not be restricted for reasons of efficiency alone. On the other hand, errors can also occur in human decision-making.

[16] European Commission, Guidelines on the application of Council Directive 2004/113/EC to insurance, in the light of the judgment of the Court of Justice of the European Union in Case C-236/09 (Test-Achats), para. 15.

[17] European Commission, Guidelines on the application of Council Directive 2004/113/EC to insurance, in the light of the judgment of the Court of Justice of the European Union in Case C-236/09 (Test-Achats), para. 14.

[18] European Commission, Guidelines on the application of Council Directive 2004/113/EC to insurance, in the light of the judgment of the Court of Justice of the European Union in Case C-236/09 (Test-Achats), para. 14.

[19] European Commission, Guidelines on the application of Council Directive 2004/113/EC to insurance, in the light of the judgment of the Court of Justice of the European Union in Case C-236/09 (Test-Achats), para. 14.

[20] European Commission, Guidelines on the application of Council Directive 2004/113/EC to insurance, in the light of the judgment of the Court of Justice of the European Union in Case C-236/09 (Test-Achats), para. 14.

Finally, the method must also be appropriate. The measure is appropriate if the disadvantages caused are not disproportionate to the purpose pursued and do not cause excessive impairment of those who invoke the protected characteristic.[21]

Within the framework of the balancing of interests, it must be examined to what extent a decision-making process, which is the same for all parties involved, can be deviated from in favor of a "better" result. Despite the process-oriented understanding of Directive 2006/54/EC, result-oriented aspects cannot be completely disregarded even if there is a legal justification. The two aspects are interrelated and influence each other. In the scope of Directive 2006/54/EC, however, it must now be noted that the process-oriented approach specified there may not be completely displaced by result-oriented arguments without further ado. Ultimately, this would circumvent the concept chosen by the legislator, who deliberately chose a process-oriented rather than a results-oriented approach. In the present case, the interests of three groups must be taken into account: The interests of male applicants who are not selected because of the threshold assigned to them. The interests of the female applicants who are taken because of the lower threshold now applicable to them. And lastly, the employer's interests in selecting the best candidates. From the perspective of the male applicants, the argument against an appropriate balancing of interests could be that they would have exceeded the threshold by using a single decision formula and would thus have been selected as suitable applicants, but are now not hired because of the separate decision formula and the higher threshold. It could be concluded that in this constellation, men are systematically discriminated against because they have to reach a higher threshold than female applicants. However, what was said earlier should be emphasized here as well: The bottom line is that men are not required to meet higher thresholds because they are men, but because the underlying data show that men below the threshold are actually less qualified for the job. By optimally assessing employees, it is clear in the fictitious scenario that different characteristics need to be considered differently for women and men in certain contexts. The employer's interest in a differentiated calculation method outweighs the interest in a uniform assessment, since there can be no interest in a uniform assessment process whose result has errors if an error-free selection method is available at the same time. The systematic discrimination of women, who are not selected even though they would have been more suitable for the job, which prevails in a joint consideration is eliminated. The result now obtained is perfect because it has no errors. Although it is primarily result-oriented aspects that weigh heavily here, ultimately there is no deviation from the process-oriented approach. Rather, the process is adapted in such a way that it delivers the best possible results. The result takes into account both individual justice and justice for the whole of society and assigns a higher value to both.

6.3. Justification in reality

Now, the consideration of this fictitious scenario is of a very theoretical nature. After all, how likely is it to find such a simple-to-analyze world in which error-free selection is possible? If the scenario is shifted to reality, the consideration becomes more complex, because the fictitious scenario leaves out some aspects that play a major role in the legal evaluation. If these aspects are included in the scenario, new risks and disadvantages can arise because our world is not that simple.

The legitimate purpose remains the best possible selection interest of the employer. In reality, however, the question arises as to what is a "better" outcome in the first place. In our fictitious scenario, it was clear that no mistakes would be made. In reality, there will be wrong decisions in any case. As long as the employer's applicant selection is more successful overall and it can be inferred that the company is performing better overall, this is a legitimate purpose [29, p.488].

However, it is questionable whether splitting the data set by gender and performing SVM analyses are also appropriate to achieve a better outcome. This starts with the construction of the SVM. If the algorithm is to find the best applicants, then a selection based solely on skills and suitability for the job in question must also take place in the training data. Frequently, however, extraneous considerations also affect the selection process. For example, the probability of not being hired is higher for pregnant women or women who wear a headscarf [30]. A data set is always a reflection of the society from which the information has been obtained. If the society contains discriminatory elements and structures, these are also present in the training data set. The SVM then analyzes a data set that is not optimal, but rather one that is loaded with prejudices. The results of the analysis are then incorporated into the

[21] CJEU, C-83/14 - CHEZ Razpredelenie Bulgaria AD, 16.7.2015, ECLI:EU:C:2015:480, para. 123.

decision formula. Thus, even the training data can lead to subsequent discrimination. It is true that the same problem arises in general when algorithms are used in decision-making processes. In our scenario, however, it must now be considered to what extent a reliable evaluation can still be derived from a possibly already influenced data set. Both questions are essential, not only because they play a crucial role for the evaluation of our scenario, but also because they affect a large number of algorithmic decisions that increasingly determine our everyday life.

Furthermore, the question arises as to what extent the division into two applicant groups, in our case men and women, is at all possible and meaningful. Such a rigid sorting fails to recognize the diversity of our society and does not do it justice. On the contrary, it will inevitably lead to wrong decisions. Whether one can circumvent this problem by dividing the data set into further groups seems doubtful. For one thing, it is not clear how many groups would be needed to do justice to all those affected. One would have to form an indefinite number of other categories in addition to the "gender" category. On the other hand, a similar problem arises when assigning certain characteristics to certain gender groups, which are supposed to apply without error to all group members and cannot be found in the comparison group. Ultimately, the differentiations are always based merely on observations and statistics, which can hardly be transferred to the general public without error.

Behind both problems lies the need to come as close as possible to case-by-case justice. This principle is not only firmly anchored in European anti-discrimination law; the European Court of Justice also applies it strictly [31, pp.29].[22]

In reality, mistakes will be made due to the actual circumstances and the technical possibilities. If it were determined that the differentiating model has a "better" overall result, the question arises as to whether more fairness for society as a whole may be achieved at the expense of fairness in individual cases. This must directly involve consideration of how much "better" an outcome must be compared to the common analysis, by what standard this should be measured, and who has the legitimacy to set that limit.[23]

Following on from these concerns, it is also necessary to consider, in real-world circumstances, whether the process-oriented approach can be constrained by departing from an evaluation process that is the same for all participants in favor of an overall "better" outcome. Unlike the fictitious example case, in the real world an unmanageable number of factors seem to influence the scenario. This leads to the interests of employers and female applicants becoming lost in the complexity of the real world situation. Ultimately, it must be stated that a result-oriented approach cannot justify the limitation of a process-oriented approach. This is because the process-oriented approach prescribed by the legislator, which envisages justice in individual cases, would then not merely be adapted, but rather disregarded.

7. Criticism of the current legal situation

We are of the opinion that the process-oriented approach of the legislator must be questioned. It is certainly justified in the context of sensitive decisions, since in these constellations the individual must be the focus of the decision. Nevertheless, it is difficult to understand why the process-oriented approach is adhered to despite increasing algorithmic decisions, when it leads to more errors in certain constellations. There can be no interest in a poorly functioning process-oriented procedure if a better and less error-prone method is available as an alternative. A process-oriented approach is only supposedly closer to individual case justice: if it is assumed that fewer errors occur when protected characteristics are included, individuals are ultimately also less likely to be assessed incorrectly. At first glance, a results-oriented approach would lead to more justice for society as a whole, but this would also have an indirect effect on individual justice. The risk of individuals being affected by an incorrect assessment would decrease. Why should individuals be entitled to a method that is on average more prone to error? Because challenges and problems in implementing transparency requirements often result in individuals being unaware of the automated decisions that affect them, better outcomes overall could strengthen their position, rather than being tied to a process-based approach that leads to more errors that are not detected then. Even if transparency can be granted, we believe that protected characteristics can be included in decision-making processes. It is required that it can be demonstrated that the process is more successful. In order to remove the suspicion of inadmissible discrimination, it must be

[22] CJEU, C-668/15 - Jyske Finans, 6.4.2017, ECLI:EU:C:2017:278 para. 32; CJEU, C-457/17 - Maniero, 15.11.2018, ECLI:EU:C:2018:912 para. 48.

[23] The same question arises with the fairness measure "conditional independence", cf. [32].

justified that there is a causal connection between the protected characteristic, the feasibility of a particular task, and the characteristic that is assessed differently.

In our opinion, the ruling of the European Court of Justice on insurance rates does not contradict this. Even though the reasons for the ruling were very brief, it can be assumed that the idea of solidarity taken up there had a significant influence on the decision. The principle of solidarity, which arises from the welfare state principle, may justify an obligation of all to the common good in certain decisions. At the same time, however, this principle may only be used where society must stand up for one another. Since this obligation to stand up for one another may well be judged differently, the legislature, which has parliamentary legitimacy, should address this issue. In certain constellations, the principle of solidarity may require mutual responsibility for reasons of the common good. This may be the case in the health sector, for example, but not in the employment context.

Finally, it is also doubtful whether the court intended its decision to provide any direction at all for algorithmic decisions. Even though a non-negligible number of AI-controlled systems were already available in 2010/2011, they have seen considerable market growth in recent years. We also doubt that the legislator had them in mind. Last but not least, the manifold criticism of the CJEU's ruling in the literature also shows that a differentiation would not be far-fetched in the insurance sector either (see for example [33] [24]; [19]).

Fairness and justice cannot be guaranteed by fundamentally ignoring protected characteristics. Such an approach would cling to a biased decision-making process and leave out a potentially less error-prone method. By using algorithmic systems, the outcome should be given great importance in the evaluation process. [1, p.778]. The legal system should not cling to something that is out of date when, at the same time, better alternatives are available for certain constellations. This is increasingly important at a time when more and more areas of our lives are influenced by algorithmic decisions. By using algorithmic systems in selection decisions, we get the opportunity to incorporate more justice into our social structures. There are a number of approaches available to do this: On the one hand, we could consider what fairness benchmarks should apply to algorithmic decisions[24], and on the other hand, we could use algorithmic systems to evaluate our human behavior and use it to draw conclusions for the future. The question of including protected characteristics is therefore only one aspect of the challenge of how we want to deal with algorithmic systems. To address this in the best possible way, those responsible should take all approaches into account.

Legislative adjustments can be made both at national and at European level. The German legislator would be free to go even further than the European legislator in its anti-discrimination protection, as this is an area of so-called minimum harmonization.[25] So far, no concrete proposals are apparent. Even the proposal for a directive COM(2008)426 which only deals with areas of application outside employment protection and defines both the concept of direct and indirect discrimination in the same manner as in the aforementioned directives, is currently not pursued further. A deviation from the process-oriented approach is therefore not discernible.

For shaping the future, it would be desirable to include a certain degree of flexibility in anti-discrimination protection. Instead of a complete abandonment of the process-oriented approach, those responsible would have to be enabled to apply the differentiating model in certain constellations and to include certain protected characteristics in the decision-making process. These protected characteristics must be specified by the legislature. The responsible party must justify why it chooses the differentiating model and why this leads to a better result in the specific case. Statistical evidence is required for the latter. At the same time, the applicability of the model may be excluded if a particularly sensitive decision is made. This is the case if legal interests protected by the CFR and the constitutions are particularly affected. Within the framework of the justification, the causal connection between the protected characteristic, the feasibility of a certain task and the characteristic that is judged differently must then be presented. If these three aspects can be presented, the suspicion of unlawful discrimination can be eliminated and a differentiating model can be applied that can contribute to more justice.

With regard to a potential amendment of the law, the advantages and disadvantages of both the process-oriented and the result-oriented approach should be considered. Since it can be assumed that the legislator was not able to consider the use of algorithmic decision systems in many drafts, the reform should be approached with an open mind under the new circumstances.

[24] for a deeper discussion, see [34] or [32].

[25] The minimum harmonization follows from Art. 6 and Recital 25 of the Directive 2000/43/EC, Art. 8 of the Directive 2000/78/EC, Art. 7 and Recital 26 of the Directive 2004/113/EC and Art. 27 of the Directive 2006/54/EC.

8. Acknowledgements

The research was performed within the project GOAL "Governance of and by algorithms" (Funding code 01IS19020; https://goal-projekt.de/en/) which is funded by the German Federal Ministry of Education and Research. The content of this paper is the sole responsibility of its authors.

Conflict of interest

The authors declare that they have no conflict of interest.

References

[1] N. Guggenberger, Neue Methodik, billige Entscheidungen und Ergebnisvorgaben, KI und Recht:MMR (2019) 777–778.
[2] S. Barocas, M. Hardt, A. Narayanan, Fairness and Machine Learning, fairmlbook.org, 2019, http://www.fairmlbook.org.
[3] A. P. Dawid, Conditional independence in statistical theory, Journal of the Royal Statistical Society: Series B (Methodological) 41 (1) (1979) 1–15.
[4] K. Zweig, Ein Algorithmus hat kein Taktgefühl: Wo künstliche Intelligenz sich irrt, warum uns das betrifft und was wir dagegen tun können, Heyne Verlag, 2019.
[5] F. S. M. Heselhaus, C. Nowak, M. Baldus, M. Breuer, T. Bruha, M. Bungenberg, W. Cremer, C. Gaitanides, A. Haratsch, Handbuch der europäischen Grundrechte, CH Beck, 2020.
[6] R. Streinz, EUV/AEUV, Vol. 3, C.H. Beck, 2018.
[7] H. D. Jarass, Charta der Grundrechte der Europäischen Union, C.H. Beck, 2021.
[8] I. Zliobaite, B. Custers, Using sensitive personal data may be necessary for avoiding discrimination in data-driven decision models, Artificial Intelligence and Law 24 (2) (2016) 183–201.
[9] A. Narayanan, V. Shmatikov, Robust de-anonymization of large sparse datasets, in: 2008 IEEE Symposium on Security and Privacy (sp 2008), IEEE, 2008, pp. 111–125.
[10] H. Zang, J. Bolot, Anonymization of location data does not work: A large-scale measurement study, in: Proceedings of the 17th annual international conference on Mobile computing and networking, 2011, pp. 145–156.
[11] J. Meyer, H. Sven, Charta der Grundrechte der Europäischen Union, Vol. 5, Nomos, 2019.
[12] M. Pechstein, C. Nowak, U. Häde, I. Band, Frankfurter Kommentar, Mohr Siebeck, 2017.
[13] D. Merten, H.-J. Papier, Handbuch der Grundrechte in Deutschland und Europa, Vol. VI/1, C.F. Müller, 2003.
[14] C. Calliess, M. Ruffert, EUV/AEUV, C.H. Beck, 2016.
[15] D. Ehlers, Grundrechtsbindung und grundrechtsschutz von unternehmen im deutschen und europäischen recht, Deutsches Verwaltungsblatt (2019) 397–406.
[16] M. Franzen, I. Gallner, H. Oetker, Kommentar zum europäischen Arbeitsrecht, Vol. 3, C.H. Beck, 2020.
[17] U. Mönnich, Unisex: Die eugh-entscheidung vom 1. 3. 2011 und die möglichen folgen, VersR (2011) 1092–1103.
[18] G. De Baere, E. Goessens, Gender differentiation in insurance contracts after the judgment in case c-236/09, association belge des consommateurs test-achats asbl v. conseil des ministres, Columbia Journal of European Law 18 (2012) 339–367.
[19] J. Lüttringhaus, Europaweit unisex-tarife für versicherungen!, EuZW (2011) 296–300.
[20] M. Wendland, Vertragsfreiheit und Vertragsgerechtigkeit, Vol. 1, Mohr Siebeck, 2019.
[21] K. Leube, Sozialversicherung in gestalt der privatversicherung - rechtliche rahmenbedingungen, NZS (2003) 449–456.
[22] E. Schanze, Injustice by generalization: notes on the test-achats decision of the european court of justice, German Law Journal 14 (2) (2013) 423–433.
[23] G. Slettvold, A critical analysis of gender discriminatory practices in insurance law in the uk - equality at all costs, UK Law Student Review 3 (1) (2015) 16–43.
[24] D. Looschelders, Aktuelle auswirkungen des eu-rechts auf das deutsche versicherungsvertragsrecht unter besonderer berücksichtigung der geschlechtsspezifischen tarifierung, VersR (2011) 421–429.
[25] H.-P. Schwintowski, Geschlechtsdiskriminierung durch risikobasierte versicherungstarife?, VersR (2011) 164–172.
[26] K. P. Purnhagen, Zum verbot der risikodifferenzierung aufgrund des geschlechts – eine lehre des eugh zur konstitutionalisierung des privatrechts am beispiel des versicherungsvertragsrechts?, EuR (2011) 690–705.
[27] M. Heese, Offene preisdiskriminierung und zivilrechtliches benachteiligungsverbot, NJW (2012) 572–577.
[28] B. Gsell, W. Krüger, S. Lorenz, C. Reymann, beck-online.GROSSKOMMENTAR, Vol. 15.09.2021, C.H. Beck, 2021.
[29] W. Däubler, Was bedeutet "diskriminierung" nach neuem recht, ZfA (2006) 479–491.
[30] D. Weichselbaumer, Discrimination against female migrants wearing headscarves, EconStor.
[31] S. Wachter, B. Mittelstadt, C. Russell, Why fairness cannot be automated: Bridging the gap between eu non-discrimination law and ai, Computer Law & Security Review 41 (2021) 105567.
[32] M. P. Hauer, J. Kevekordes, M. A. Haeri, Legal perspective on possible fairness measures–a legal discussion using the example of hiring decisions, Computer Law & Security Review 42 (2021) 105583.
[33] C. Armbrüster, Schlussanträge der generalanwältin juliane kokott vom 30. 9. 2010 in der rechtssache c-236/09 (test-achats) zur frage der durch art. 5 abs. 2 der richtlinie 2004/113/eg (gender-richtlinie) eröffneten zulässigkeit geschlechtsbezogener differenzierungen bei versicherungsverträgen., VersR (2010) 1571–1583.
[34] S. Wachter, B. Mittelstadt, C. Russell, Bias preservation in machine learning: the legality of fairness metrics under eu non-discrimination law, West Virginia Law Review, Forthcoming.