Fairness in statistical jobseeker profiling

28.01.2021, Martin Gasser (margasser@student.ethz.ch)

Outline

For public employment services, the idea to use statistical tools to profile jobseekers is appealing and such tools have found widespread use in OECD countries. They promise a more efficient and effective allocation of Active Labour Market Programmes (ALMP). Although the Swiss Unemployment Insurance (ALV) currently uses no statistical jobseeker profiling, it is likely only a matter of time before this matter will have to be considered.

When designing a statistical profiling, careful consideration should be given to to potentially harmful and adverse effects. To help give some structure to such an assessment, I propose four guidelines for the ALV. These guidelines are intended as a starting point for discussion; to be complemented by legal requirements and further standards at a later stage. The first guideline is mainly procedural; the remaining three are more substantive.

- Guideline 1: Provide an analysis of potential harms and potential adverse effects. This should happen at the outset of algorithm development and include important stakeholders.
- Guideline 2: In order to have clear accountability, prediction and decision should be separated. Developers should be responsible for prediction quality; the ALV should be responsible that there is no discrimination in its profiling-based allocation of ALMPs.
- Guideline 3: Some specific forms of discrimination can be measured by formal criteria. An appropriate default criterion is proposed: the criterion of "equalised odds". It measures whether certain protected groups are equally affected by classification errors (false positives and false negatives). This can also be enforced by using different group-specific thresholds (at the cost of some prediction accuracy).
- Guideline 4: Potential discrimination should not only be discussed separately for each legally protected attribute (e.g. age, gender), but also with respect to particularly vulnerable intersectional groups that rely more heavily on having access to the right kind of ALMPs.

The case of jobseeker profiling

In order to obtain unemployment benefits, jobseekers have to register at the public employment service. They are required to disclose a broad range of personal data to verify benefit claims and for job matching. This includes, for instance, information on education, employment history, language skills and previous income. The broad range of administrative data makes the idea of a statistical jobseeker profiling obvious.

Usually, statistical profiling tools estimate the risk of long-term unemployment in order to help targeting Active Labour Market Programmes (ALMPs). The term AMLP is understood broadly to comprise all actions taken by public employment services, for example job referrals, counselling by caseworkers, job training, temporary employment programmes or subsidised internships. As shown in the two figures below, from the World Bank [1] and from the OECD [2], statistical jobseeker profiling usually produces a risk score, based on which jobseekers are then categorised into low, middle or high risk groups. Decisions remain at the discretion of caseworkers and counsellors, but are informed by the profiling output. The decision making is thus at most semi-automated. The exact purpose and design of profiling tools vary in practice [3].

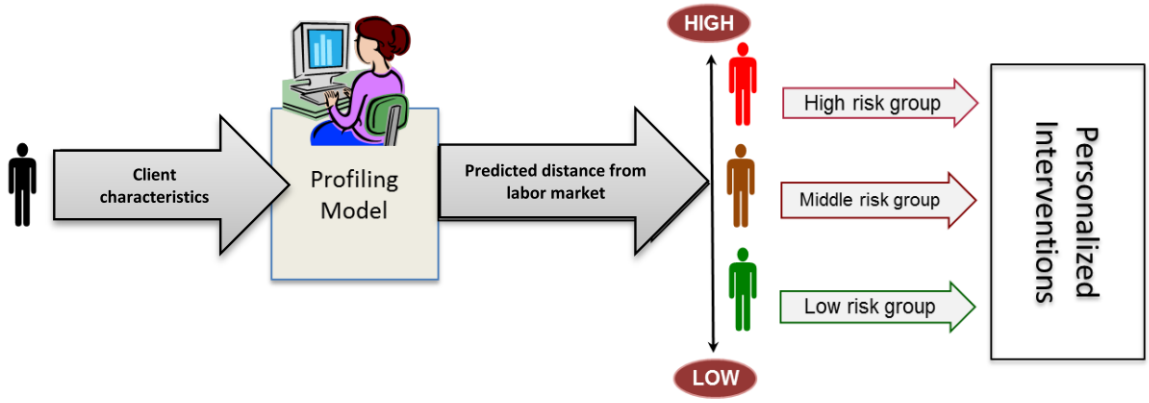
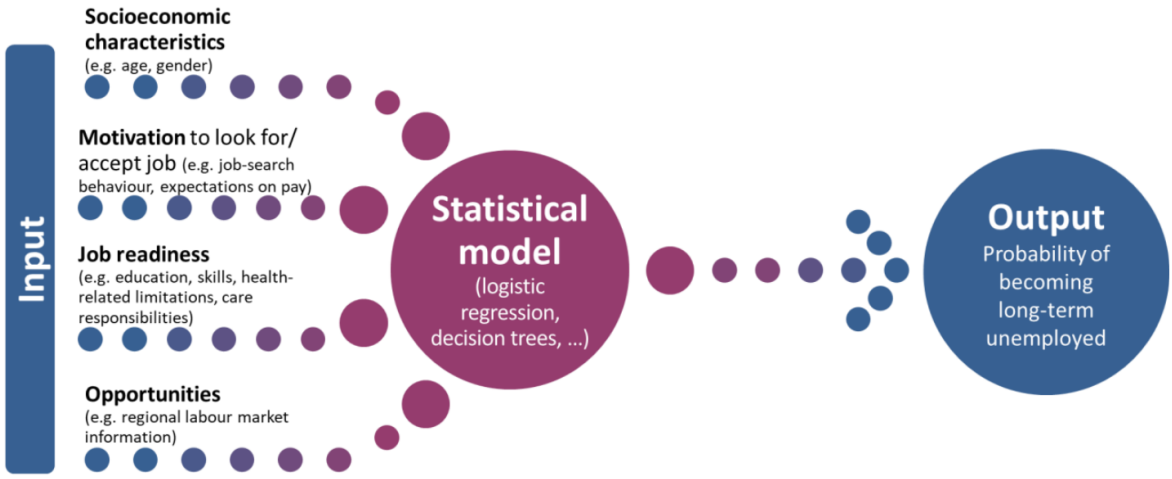


Figure 1: Figure from a World Bank Group report on profiling the unemployed [1].

Figure 2. The building blocks of statistical profiling models



Source: Authors.

Figure 2: Figure from an OECD report on profiling in public employment services [2].

The argument for statistical jobseeker profiling is appealing. Often, difficulties in finding a job become apparent only too late, and "negative duration dependence" means that the callback rate for job interviews declines during the early stage of unemployment (e.g. [4]). Useful and supportive ALMPs are thus either deployed too late, partly assigned to people who do not need them (deadweight loss), or become punitive if the people are sent to the wrong programmes. Generally speaking, profiling is seen as a desirable step towards individualised service: "*As a result, statistical profiling gives counselors an objective basis for differentiating intervention intensity [...] Statistical profiling enables [public employment services] to deploy special interventions for the high-risk cases before [long-term unemplyoment] manifests*" ([1], p. 18-19).

But statistical jobseeker profiling also raises important fairness concerns. This was recently experienced in Austria, where such a tool (called AMAS) was developed to categorise jobseekers into the low, middle and high risk groups. The idea was to allocate ALMPs primarily to the middle group [5]. It then emerged that AMAS gave women lower average scores, especially if they also had childcare obligations (while men did not face such a caregiver penalty) or a migrational background (which also had no effect on men) [6]. The model drew intense criticism and legal challenges, it ultimately required a decision by the highest administrative court for AMAS to be allowed [7].

These considerations suggest, that the Swiss unemployment insurance (ALV) should prepare guidelines for the development and use of such profiling tools. No such tool is currently in use, but this is probably a matter of time and guidelines should be available in the run-up to the development.

The setup under consideration is thus: Based on personal data X (self-reported data as well as administrative records) and protected individual attributes A (e.g. gender), the statistical model predicts the probability of long-term unemployment Y: $\hat{\pi} = \hat{P}(Y = 1|X, A)$. Long-term unemployment is defined as 12 or more consecutive months of registered unemployment. [1] The actual profiling R takes one of the three values "high risk", "middle risk" or "low risk" and is achieved by thresholding the probability: $R = h(\hat{\pi})$. For example, the Austrian system AMAS classifies jobseekers as "low risk" if they had a at least a two-thirds chance of finding work within half a year, and as "high risk" if they had at least 25% chance of not finding work within two years. The classification R is then used to inform the allocation of ALMPs. A likely choice is to focus ALMPs on the middle-risk segment, which is assumed to benefit from supportive ALMPs, but without requiring the kind of specialized support not available to the ALV but needed for jobseekers with multiple barriers to employment.

The following guidelines assume that a reasonably precise profiling can be developed. This is not obvious and at least one previous attempt failed [8]. Finding a job is inherently stochastic with a good deal of luck involved, and extant administrative records often fail to include relevant predictors (e.g. motivation or health indicators).

These guidelines are intended as a starting point for discussions within the ALV. They are intended to be part of a larger (yet to be developed) catalogue of requirements (e.g. transparency, interpretability of the classifier, explainability of individual predictions).

Guideline 1: Analyse potential harms and adverse effects at the outset

In the run-up to the development of a statistical profiling tool, the developers are required to present and publish an analysis of potential harms ("Technik- und Risikofolgenabschätzung"). This includes working with the ALV and practitioners (e.g. caseworkers) to identify use cases and potential sources of harm. There is a variety of templates and examples for such analyses [e.g. 10, 23, 24].

The analysis of potential harms should treat statistical jobseeker profiling as a fully-automated decision, even though it is intended to only provide input for caseworkers. In practice, the degree of caseworker discretion when using the profiling results will be under constant pressure (ch. 2.4. in [9]). First, in order to make good on the promise of efficiency gains, caseworkers have to routinely accept the profiling results. Otherwise they would spend as much time on assessing each case as without a tool. Second, to deliver on the promise of objectivity and prevention of arbitrariness, caseworkers should accept the profiling regularly. Third, profiling results will routinely be adopted given that caseworkers are often short on time and have a high caseload ([5], p. 2). Fourth, there is likely a justification asymmetry. If people complain or if supervisors inquire, it might be easier to justify acting on the profiling result than deviating from it.

An incomplete list of potential fairness-relevant points to discuss includes:

- **Proportionality of data collection.** From a fairness perspective it is important to note that available data for profiling are de facto mandatory. Refusing to disclose data, e.g. about previous employment, could lead to the loss of unemployment benefits. This is especially important because statistical systems tend to raise the "appetite" for even more data: It is usually easy to think of important yet unobserved variables that might improve predictions.

This creates an impulse to collect more information, often about sensitive issues like aspirations, motivations, mental health, substance abuse, ability to work in teams etc. Making such sensitive information mandatory only to make a statistical profiling more accurate seems, *prima facie*, to place an unproportional burden on privacy. At the very least the trade-off has to be carefully argued.

- **Bias reflected in input data**. Any statistical profiling will depend on data about past unemployment spells. As a consequence, pre-existing bias and unfairness in job finding rates will directly influence the profiling. For instance, ageist stereotypes may be one reason for the difficulties that elderly jobseekers face. The observed long-term unemployment then already contains discriminatory practices.
  It is worth pointing out that our preferred measure of discrimination (see guideline 3) actually depends on the assumption that sensitive attributes can be used in profiling as long as they only affect the profiling R by affecting long-term unemployment Y. If age correlates with long-term unemployment then it is allowed to affect the profiling result. This should not be taken to mean that the effect of age on unemployment duration is fair or non-discriminatory. But discrimination on the labour market is another matter; here I am only concerned with an unfair or discriminatory impact of statistical profiling on the allocation of ALMPs.
- **Quality of input data.** Although administrative records are used, at least two important problems with data quality remain:
  - Comparability: Public employment services in Switzerland are operated by the Cantons, which have substantial freedom to adapt practices and processes. The result are data that sometimes look nominally the same for all Cantons (they use the same software and hence the same data fields), but have very different meanings. Working closely with caseworkers and practitioners could help identify some problems with the input data (an innovative case is [26]).
  - Missing data: Even data that seem very safe, e.g. financial data that have to be verified with copies of previous employment contracts, can be problematic due to non-random missingness. The reason is that at prediction time (e.g. in the first meeting between jobseeker and caseworker) the application and the necessary documents might still be under consideration. For example, complicated and non-standard work arrangements, or incompletely submitted documents due to language difficulties, mean that checking the application takes longer and the data are more likely missing at prediction time.
- **Representational harm.** Classifying jobseekers into three segments labelled "low risk", "middle risk" and "high risk" [2], might itself be potentially harmful. It suggests that risk of long-term unemployment is somehow a fixed trait of a person, instead of a complex interplay between labour market demand, professional history, personal match, luck and so on. Disproportionately classifying one group as "high risk" can therefore be stigmatizing. To mitigate the harm, profiling should be as dynamic as possible, e.g. updating the prediction over the course of the unemployment spell, as well as taking into account developments on the labour markets. This could avoid reading the scores as personal traits and stresses their dynamic and contextual nature.
- **Feedback effects on jobseekers**. Related to the representational harm is the danger of jobseekers being discouraged by a "high risk" classification, or feeling overly optimistic due to a "low risk" profile. Another type of feedback loop is strategic behaviour, if ALMPs are tied to certain categories. If only "high risk" jobseekers are eligible for income subsidies, then there is an incentive to "game the system" by (mis-)reporting data so as to be classified "high risk". This would, in turn, deteriorate the input data for the statistical profiling.
- **Effects on caseworkers**. Evidently, a statistical profiling tool has substantial consequences for caseworkers who would use it every day.

    - The most obvious is that it alters their job and raises fears of "digitization" or of being "replaced by a computer". Two pilot projects in Switzerland (one a statistical jobseeker profiling, one a econometric targeting system for ALMPs) failed partly due to resistance by caseworkers to actually use the system and follow its recommendations ([8, 22]). Giving a voice to caseworkers and other practictioners in all development and implementation steps will be necessary.
    - A second danger is linked to the above-mentioned "appetite" for ever more data. For caseworkers, this often means having to verify more and more data. This entails a lot of going through checklists and following prompts from the tool, thus crowding out time for human interaction and counselling. "*This changes the working practice of AMS counsellors significantly from personal counselling to the processing of a series of "tasks" as required by the assistance system*" [5].
    - A third effect is "gaming" the profiling tool by caseworkers. If they consider an ALMP especially useful for a jobseeker, and this ALMP is to be used only for the segment "middle risk", she might try to "bend" the data in direction of this segment (note that often information about jobseekers is gathered or at least double-checked by the caseworker). This will generally lead to less reliable data.
    - A fourth harmful effect could be "Stereotype reinforcement" and "Confirmation bias" [10]: The tendency to discount (or explain away) output from the profiling tool that does not match stereotypes while interpreting any stereotype-matching output as an "objective" verification.
- **Dignitary considerations**. Although this is rather vague, statistical profiling also raises important questions relating to personal dignity: Do we really prefer a labour-market policy that is more mechanical to a more personal but less efficient one? Is there not a danger of "reducing personal biographies to computer-generated values"? As argue [5, p. 2]; "*The shift from personal needs to population-based calculations has also serious consequences for jobseekers. For them, the focus on the [predicted] value and the classification means that their biography and abilities are reduced to a seemingly 'objective' value that is supposed to provide information about their prospects on the labour market.*"

Guideline 2: Separate prediction and decision

An essential legal constraint on any jobseeker profiling is non-discrimination, as stated in Art. 8 of the Swiss constitution [11]. Some aspects of discrimination are measurable with formal criteria, which raises two questions: Who should be held responsible if the formal criterion indicates discrimination (guideline 2) and which such criterion should be used (guideline 3)?

Discrimination with respect to jobseeker profiling is distinct from dicrimination on the labour market: It refers not to recruitment decisions by employers but to actions taken by the ALV based on the profiling output R. In this narrow sense, discrimination can be reduced by adjusting how R depends on the output of the profiling tool (i.e. the predicted probability $\hat{\pi}$ of long-term unemployment). In the fairML-literature this is called a post-processing method, because it does not change the predictor $\hat{\pi}$. Alternatives would be altering data before training (pre-processing) or penalizing the empirical loss function with a measure of discrimination (in-processing) (see e.g. [12, 15]).

Post-processing methods make separating prediction and decision easy: Developers of the profiling tool deal with the technical issue of providing the best possible prediction $\hat{\pi}$, the ALV then decides how R depends on $\hat{\pi}$ and the (cantonal) public employment services ultimately decide how R informs the allocation of ALMPs (they only ever see R, not the underlying $\hat{\pi}$). This has several advantages:

- Accountability: The responsibility for the non-discriminatory use of jobseeker profiling should rest mainly with the ALV, not with scientists or engineers developing the profiling tool. Holding the ALV responsible for the use of predictive scores automatically involves the relevant oversight bodies and political stakeholders. (Engineers are still required in the first guideline to map the potential harms and work with the ALV to find ways to mitigate them.)
- Feasibility studies: The difficulty of predicting long-term unemployment means that a first step will always be a feasibility study. One such study in the Canton Neuchâtel is currently underway [27]. It is unrealistic to assume, as pre- and in-processing approaches do, that the debate around non-discrimination has concluded already before development begins. Usually such debates only become heated once the system has been introduced and actual cases emerge.
- Simplicity: Separating prediction from decision makes non-discrimination easier to understand and discuss for non-technical stakeholders. They do not have to understand *"a possibly complex training pipeline"* ([14], p. 3), such as penalization of discrimination in the training objective. Nor do they have to accept discrimination if an algorithm trades off better accuracy for more discrimination (in-processing). The separation also makes it much easier to change thresholds after development of the profiling tool.
- Theoretical reasons: Especially costly types of ALMP are limited to only a small number of jobseekers. In such a setting, the optimal way to proceed is to take the best available predictive scores, then using different cutoffs in decision making [13, 21]. "*Equity preferences can change how the estimated prediction function is used (such as setting a different threshold for different groups) but the estimated prediction function itself should not change.*" ([13, p. 23]). In addition, early empirical evidence indicates that post-processing via different thresholds works well on real problems, decreasing accuracy less than competing approaches [12].

Guideline 3: Use equalised odds as an imperfect diagnostic criterion

Many formal criteria of non-discrimination have been proposed in the literature [15]. Some are observational in the sense that they only depend on the values of $\hat{\pi}$, R, A and Y. It is well-documented in the literature that some of the most common criteria are mutually incompatible, i.e. they cannot be fulfilled at the same time [15]. For example, if two groups have different base rates of long-term unemployment, the prediction cannot be equally well-calibrated for both groups and have equal error rates (false positive and false negative rates) for both groups [20]. This suggests that the ALV will at some point have to prioritize one criterion over the others.

In the context of statistical jobseeker profiling the criterion called "equalised odds" seems appropriate.[3] Technically speaking, it measures whether the classification R is statistically independent of the protected attributes A, given the true outcome Y [14].[4] The predicted score is allowed to depend on attributes A, but only insofar as A correlates with the true outcome. For example, newly registered jobseekers aged 55+ have a 40% chance of long-term unemployment, whereas those aged 25 to 35 have only a risk of 20% (p. 31 in [16]). Clearly, older jobseekers will more often be classified as “higher risk” than younger ones, which does not violate the equalised odds criterion, as long those older jobseekers who actually go on to become long-term unemployed are not more often wrongly classified as “middle or low risk”. Having equalised odds implies that the false positive rates and the false negative rates are equal, or almost equal, for the different groups in A [15].

The criterion of equalised odds thus measures whether the protected groups have the frequency of being wrongly classified "low risk". And the same holds for being wrongly classified "middle risk" and being wrongly classified "high risk". It thus treats the different kinds of errors as "incomparable" in the sense that more false positives cannot be balanced by less false negatives

or vice versa. Alternative formal non-discrimination criteria (e.g. calibration or statistical parity) fail to ensure that false positives and false negatives occur at similar rates across all protected groups; they are willing to trade-off false positives with false negatives. This is what happened in the famous COMPAS scores of recidivism, where one group was primarily affected by false positives, the other group primarily by false negatives [15].

Not trading-off different kinds of error is important in our context because labour-market policy in Switzerland is highly decentralized and the allocation of ALMPs based on the profiling output depends on Cantonal policies. One Canton might use a score to assign a supportive and very costly measure to high risk jobseekers; another Canton might use the same score to assign milder supportive measures to the broad group of low or middle risk jobseekers. And the same ALMP such as temporary employment can be beneficial for some jobseekers but feel degrading and punitive to others. This complexity means that it does not make sense to generally that one type of error is more important than the other. The criterion of equalised odds has no need to do this: It ensures that both kinds of error (false negatives and false positives) are equally likely across the different protected goups. Another reason for treating the two kinds of error as incomparable is that they can be qualitatively different. If a jobseeker with a low risk of long-term unemployment is wrongly assigned to a an intensive training program based on the output "high risk", the costs are borne by the ALV. But if a high-risk jobseeker cannot participate in the training program because she is wrongly assigned a low risk score, it is mainly her who bears the cost of the wrong prediction.

As already mentioned, equalised odds is an observable criterion and can be measured and reported. Unequal odds can then be taken as indicating that somewhere in the modelling process (e.g. collecting training data, "cleaning" data, updating the model, the algorithm itself etc.) there is a bias that results in discrimination [17]. The criterion can thus be used as a diagnostic test for a very specific form of discrimination and can inform a discussion about the use of statistical profiling. However, given a predicted probability $\hat{\pi}$ from the profiling, equalised odds can also be enforced by choosing group-specific thresholds, although a small amount or randomization might be necessary [14]. But it should be kept in mind that doing so is by no means a guarantee for non-discrimination more broadly construed [15].

Although equalised odds seems an appropriate criterion in the context of statistical jobseeker profiling, this remains something to be discussed with the relevant stakeholders. Given the specific circumstances that arise with a concrete tool, they may also opt for another criterion of non-discrimination instead.

Guideline 4: Give special consideration to vulnerable groups

The Swiss constitution provides a (non-exhaustive) list of protected groups in Art. 8 No. 2: "*No person may be discriminated against, in particular on grounds of origin, race, gender, age, language, social position, way of life, religious, ideological, or political convictions, or because of a physical, mental or psychological disability.*"[25]. This does not preclude that these attributes are used in decisions, but their use has to be particularly qualified (p. 39, [11]).

Of those protected attributes, administrative data in the context of jobseeker profiling is available on gender, age, language and country of origin. The selection and definition of protected groups should be discussed transparently with stakeholders before implementing any profiling tool.

In principle, discrimination could be discussed for each of the protected groups independently of the others. But this approach neglects the potential compounding effect of unequal access to ALMPs; it neglects intersectionality in the sense of "intercategorical complexity" [18]. Guideline 4 thus requires us to carefully think about how statistical profiling affects vulnerable groups with regard to access to ALMPs. For example, perceived "social distance" (measured via language skills and country of origin) decreases the likelihood of being assigned to ALMPs intended for upskilling [28]. Furthermore, there is some evidence that low-skilled jobseekers and migrant jobseekers are less frequently assigned to certain ALMPs [29]. Other groups may also face a potential "access bias" with regard to ALMPs, including for example migrant women and mothers with childcare responsibilities [6].

Considerations of intersectionality suggest that all intersections should be treated as protected groups. However, this would lead to a combinatorial explosion in the number of protected groups [5], thus decreasing prediction accuracy and ultimately introducing more "artificial" randomness into the classification. [6] If error rates are to be equal across all groups, then the maximum performance will naturally be upper-bounded by the group with the worst performance [14]. It would thus be preferable to apply the criterion of equalised odds to each of the protected groups separately. Although mathematically this does not guarantee equalised odds for all intersectional groups, it may be enough in practice: It seems difficult to construct cases where error rates differ substantially between intersections but not for the main groups. [7]

Setting aside such technical issues, intersectionality could instead be interpreted in a more qualitative way, as challenging us to consider carefully which specific groups may be subject to access bias or may be more reliant on ALMPs. This is indeed a crucial step in the analysis of potential harms. A clear definition of protected groups also creates the right incentive: Knowing from the start for which groups equalised odds will be measured, developers have to make sure that the satistical profiling performs well in those groups. Nonetheless, because such groups may be small, predictive accuracy could still suffer if we require equalised odds. As an non-technical alternative, particularly vulnerable groups could be excluded from the equalised odds criterion but explicitly recompensated, e.g. by granting additional access to desirable and useful ALMPs.

Conclusions

In this paper, I have discussed a concrete application of semi-automated decision making, namely the potential use of statistical jobseeker profiling by the Swiss unemployment insurance. Four guidelines are proposed in order to help structure the assessment of potentially harmful and adverse effects of such a profiling tool. These guidelines are intended as a starting point for discussion; they are also meant to be complemented by legal requirements and further standards.

The first guideline is procedural and probably the most important. It requires an analysis of potential harms early in the process, thus forcing stakeholders to define clear use cases, critically assess possibly harmful effects, and work closely with practitioners. The discussion provides a few hints at what kind of harms should be considered.

The three remaining guidelines are more substantive and provide input for the discussion of potential harms. The second guideline proposes that the profiling should be built in such a way that developers are responsible for predictive accuracy, while the ALV is responsible for the non-discriminatory use of these predictions. The third guideline proposes, as a default, using the well-known formal criterion of equalised odds to measure discrimination. The fourth guideline makes it necessary to define groups that are especially reliant on having access to the right ALMPs and thus merit special attention in the discussion of discrimination and potential harms.

As a limitation I should mention that most of the issues discussed above, as well as potential solutions, relate the literature on "fairness in machine learning". This is a very active area of research; and many important questions are still contested and open to debate. The proposed guidelines and recommendations will thus have to adapt to an eventually emerging consensus about best-practices and standards.

References

[1] Loxha, Artan and Matteo Morgandi (2014). Profiling the Unemployed: A Review of OECD Experiences and Implications for Emerging Economies. *Social Protection & Labor Discussion Paper* No. 1424. World Bank Group.

[2] Desiere, Sam, Langenbucher, Kristine and Ludo Struyven (2019). Statistical profiling in public employment services: An international comparison.* OECD Social, Employment and Migration Working Papers* No. 224. https://dx.doi.org/10.1787/b5e5f16e-en

[3] Barnes, Sally-Anne and Sally Wright (2015). Identification of latest trends and current developments in methods to profile jobseekers in European Public Employment Services: Final report. European Commission Directorate-General for Employment, Social Affairs and Inclusion.

[4] Kroft, Kory, Lange, Fabian and Matthew J. Notowidigdo, Duration Dependence and Labor Market Conditions: Evidence from a Field Experiment. *The Quarterly Journal of Economics*, Volume 128, Issue 3, August 2013, Pages 1123–1167, https://doi.org/10.1093/qje/qjt015

[5] Allhuter, Doris and Astrid Mager (2021): How fair is the AMS algorithm? ITA Dossier. http://epub.oeaw.ac.at/ita/ita-dossiers/ita-dossier052en.pdf

[6] Allhutter, Doris, et al. (2020): Algorithmic Profiling of Job Seekers in Austria: How Austerity Politics Are Made Effective. https://www.frontiersin.org/articles/10.3389/fdata.2020.00005/full

[7] Bundesverwaltungsgericht (2020): Decicion W256 2235360-1, https://rdb.manz.at/document/ris.bvwg.BVWGT_20201218_W256_2235360_1_00

[8] Arni, Patrick and Amelie Schiprowski (2015). Die Rolle von Erwartungshaltungen in der Stellensuche und der RAV-Beratung - Teilprojekt 2: Pilotprojekt Jobchancen-Barometer: Erwartungshaltungen der Personalberatenden, Prognosen der Arbeitslosendauern und deren Auswirkungen auf die Beratungspraxis und den Erfolg der Stellensuche. *IZA Research Report* No. 70.

[9] Kollek, Alma and Carsten Orwat (2020). Mögliche Diskriminierung durch algorithmische Entscheidungssysteme und maschinelles Lernen – ein Überblick. Büro für Technikfolgenabschätzung beim Deutschen Bundestag. https://publikationen.bibliothek.kit.edu/1000127166

[10] https://fpf.org/blog/unfairness-by-algorithm-distilling-the-harms-of-automated-decision-making/

[11] Staatskanzlei Kanton Zürich (2021). Einsatz Künstlicher Intelligenz in der Verwaltung: rechtliche und ethische Fragen. https://www.zh.ch/content/dam/zhweb/bilder-dokumente/themen/politik-staat/kanton/digitale-verwaltung-und-e-government/projekte_digitale_transformation/ki_einsatz_in_der_verwaltung_2021.pdf

[12] Lamba, Hemank, Rodolfa, Kit and Ghani Rayid (2021). An Empirical Comparison of Bias Reduction Methods on Real-World Problems in High-Stakes Policy Settings. *ACM SIGKDD Explorations Newsletter*, *23*(1), 69-85

[13] Kleinberg, Jon, Ludwig, Jens, Mullainathan, Sendhil and Ashesh Rambachan (2018). Algorithmic Fairness. *AEA Papers and Proceedings*, 108, 22-27.

[14] Hardt, Moritz, Price, Eric and Nathan Srebro (2016). Equality of Opportunity in Supervised Learning. https://arxiv.org/abs/1610.02413v1

[15] Barocas, Solon, Hardt, Moritz and Arvind Narayanan (2019). Fairness and Machine Learning. http://www.fairmlbook.org

[16] Egger, Dreher und Partner and Ecoplan (2020). Langzeitarbeitslosigkeit – Hürden der Arbeitsmarktintegration und Massnahmen der Regionalen Arbeitsvermittlungszentren. *SECO Publikation Arbeitsmarktpolitik* No 58.

[17] Hellman, Deborah (2020). Measuring algorithmic fairness. *Virginia Law Review*, 106, 811.

[18] Grimmett, Geoffrey and David Stirzaker (2020). Probability and random processes. Oxford University Press.

[19] McCall, Leslie (2005). The complexity of intersectionality. *Signs: Journal of women in culture and society*, 30(3), 1771-1800

[20] Chouldechova, Alexandra (2017). Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big data* 5.2: 153-163.

[21] Corbett-Davies, Sam et al. (2017). Algorithmic decision making and the cost of fairness. In *Proceedings of the 23rd acm sigkdd international conference on knowledge discovery and data mining*, 797-806.

[22] Behncke, Stefanie, Frölich, Markus and Michael Lechner (2007). Abschlussbericht zum Pilotprojekt Statistisch assistierte Programmselektion (SAPS). *Schweizerisches Institut für Aussenwirtschaft und Angewandte Wirtschaftsforschung (SIAW).*

[23] [9] Allhutter, D., Mager, A., Cech, F., Fischer, F., & Grill, G. (2020). *Der AMS Algorithmus - Eine Soziotechnische Analyse des Arbeitsmarktchancen-Assistenz-Systems (AMAS)*. doi:/10.1553/ITA-pb-2020-02

[24] https://algorules.org/de/startseite and https://algorules.org/fileadmin/files/alg/Handreichung_fuer_die_digitale_Verwaltung_Algo.Rules_12_2020.pdf

[25] https://www.fedlex.admin.ch/eli/cc/1999/404/en

[26] Gennatas, Efstathios et al. (2020): Expert-augmented Machine Learning. In *Proceedings of the National Academy of Sciences*, 117 (9) 4571-4577 https://www.pnas.org/content/117/9/4571.

[27] https://wp.unil.ch/govai/algorithmic-profiling/

[28] Auer, Daniel, and Flavia Fossati. "Compensation or competition: Bias in immigrants' access to active labour market measures." *Social Policy & Administration* 54.3 (2020): 390-409.

[29] Bonoli, Giuliano, and Fabienne Liechti. "Good intentions and Matthew effects: access biases in participation in active labour market policies." *Journal of European Public Policy* 25.6 (2018): 894-911.

---

1. Because this outcome (long-term unemployment) covers *registered* unemployment only, there are by definition no missing values. Such a narrow focus can be justified in this case, as only people registered as unemployed can have access to ALMPs and the profiling tool is intended to inform the allocation of ALMPs. ↩

2. This is the OECD terminology. In the Swiss version they are "leicht vermittelbar" (easy to place), "mittel vermittelbar" and "schwer vermittelbar"; there is no mention of "risk". ↩

3. This criterion goes by several names, e.g. "separation" or "predictive equality". ↩

4. Formally, this reads as: $R \perp A|Y$. The output R takes the three values "low risk", "middle risk" and "high risk" and is thus not binary. The criterion needs to cover cases where an ALMP is targeted to one of the three groups. Equalised odds holds for all possible such combinations if we simply require that $\hat{\pi} \perp A|Y$. Since $R = h(\hat{\pi})$ (h being a thresholding or step-function) if follows that $R \perp A|Y$ (e.g. p. 49 in [18]). ↩

5. As an example of the combinatorial explosion, consider just two origin groups, two gender groups, two age groups and three language groups we already get 24 groups: $A = (A_1, A_2, A_3, A_4) \in \{0,1\} \times \{0,1\} \times \{0,1\} \times \{0,1,2\}$. Requiring equalised odds for those 24 groups is strictly more demanding than requiring it for each of the protected attributes separately. ↩

6. We can calculate for each group its own ROC curve [14]. These curves relate the false positive rate on the x-axis with (1 - false negative rate) on the y-axis. In other words, the group-specific ROC curves contain all the information to assess equalised odds. Feasible for overall prediction accuracy are only points below all the group-specific ROC curves. For certain groups this might mean moving to a point below the curve, which can be done by introducing randomization into the thresholding function h in $R = h(\hat{\pi})$. Concretely, we would use a lower threhold τ_l with probability p and a higher threshold τ_u with probability $1 - p$. All this, however, comes at the cost of accuracy. ↩

7. I constructed (hypothetical) confusion matrices for 2000 jobseekers in a simple case of binary classification (high risk versus low or middle risk) and a binary outcome (long-term unemployed yes/no). The classifier has an accuracy of 72.8%. Equalised odds is satisfied when comparing men and women: False positive rates (FPR) are 31.3% and 31.0%; false negative rates (FNR) are 16.2% and 16.3%. It is also satisfied when comparing Swiss to migrant jobseekers: FPR are 31.1% and 31.3%; FNR are 16.3% and 16.2%. But whereas the FNR for Swiss women is below 1%, it is 25% for migrant women. However, this example is somewhat contrived and was not easy to construct, which indicates that such a situation may be rare in practice. ↩