Proceedings of the Weizenbaum Conference 2023:

AI, Big Data, Social Media, and People on the Move

WHY DOES THE AI SAY THAT I AM TOO FAR AWAY FROM THE JOB MARKET?

Berman, Alexander
Centre for Linguistic Theory and Studies in Probability (CLASP)
Department of Philosophy, Linguistics and Theory of Science
University of Gothenburg
Gothenburg, Sweden
alexander.berman@gu.se

KEYWORDS

public employment services; decision-support systems; explainability; machine learning; artificial intelligence; explainable artificial intelligence

DOI: 10.34669/wi.cp/5.2

1 INTRODUCTION

As artificial intelligence (AI) is increasingly being deployed in various domains such as healthcare (Qayyum et al., 2021), finance (Dastile, Celik & Potsane, 2020) and public welfare (Saxena et al., 2020; Carney, 2020), there is a growing need for understanding how stakeholders are affected by AI (Vaassen, 2022) and how to design and present explanations of AI-based decisions in ways that humans can understand and use (Miller, 2019). This paper contributes to these efforts by examining an AI-based decision-support system (DSS) launched by the Swedish Public Employment Service (PES) in 2020. Specifically, the study investigates to what extent the studied system enables affected jobseekers to understand the basis of AI-assisted decisions, to negotiate or contest dispreferred decisions, and to use the AI as a tool for increasing their job chances.

The rest of the paper is organised as follows: Section 2 situates the study in relation to previous work. Section 3 presents the empirical material, including technical information about the studied DSS. The main contribution of the paper is then presented in section 4 which elaborates weaknesses and limitations in the explainability of the system and how they could be addressed. Finally, section 5 offers some conclusions.

2 RELATED WORK

Previous studies have investigated the use of AI and algorithms in the context of PES from the perspectives of accuracy and discrimination (Desiere, Langenbucher & Struyven, 2019; Desiere & Struyven, 2021), norms and values embedded in algorithms (Sztandar-Sztanderska & Zielenska, 2020), austerity politics (Allhutter et al., 2020), caseworkers' attitudes and strategies (Assadi & Lundin, 2018; Sztandar-Sztanderska and Zielenska, 2022) and legal certainty (Carlsson, forthcoming). Few previous works have analysed explainability in relation to PES; exceptions include Niklas et al. (2015) who investigated the transparency of a Polish algorithmic profiling system and Zejnilovic et al. (2021) who studied the effects of explanations on caseworkers' decisions. An important basis for the present study is Scott et al.'s (2022) investigation of jobseekers' needs and desires in relation to algorithmic DSS. This paper extends previous work by technically describing the new Swedish AI-based DSS and by analysing explainability from the perspective of jobseekers' needs and interests.

3 CASE DESCRIPTION

The material presented below is based on public sources (cited where relevant) and information received from the agency via email (Nov 2021 – May 2023).

3.1 General information

In 2019, the Swedish government decided that a statistical tool[1] should be developed as an integrated part of the operations of the Public Employment Service (PES) in order to improve consistency and accuracy of labour-market related assessments, and thereby improve efficiency of resource

[1] The term "statistical assessment support tool" is used by both the government and the Swedish PES. In this paper, the terms "AI-based" and "statistical" are used interchangeably.

allocation.[2] Subsequently, the employment initiative Prepare and Match was launched in 2020 and rolled out nationally in 2021 (Hansson et al., 2022). The initiative enables enrolled jobseekers to get support, e.g. in the form of training or guidance, from a chosen provider. Decisions about access to the initiative are based on outputs from an AI-based DSS. The function of the AI is to assess the jobseeker's distance to the job market, with the purpose of targeting the employment agency's resources to those individuals that are most likely to find a job through the initiative.

Caseworkers are instructed to primarily adhere to the automated recommendation. Overruling a negative decision is difficult since it requires contacting a special working group within the agency. Interviews with caseworkers have indicated that some of them are reluctant to use this option since the working group rarely admits exceptions from automated recommendations (Bennmarker et al., 2021).[3]

3.2 Statistical model and decision algorithm

Decisions about access to the employment initiative are partly based on a statistical estimate of the jobseeker's probability of finding a job within 6 months. The statistical analysis encompasses 26 variables pertaining to personal information, including age, gender and education, as well as previous unemployment activities. It also involves data about the jobseeker's postal area, including levels of unemployment, income, education and citizenship (Bennmarker et al., 2021).

The statistical model is a neural network[4] trained on historical data consisting of 1.1 million profiles collected over a period of 10 years. The model estimates probabilities for 14 different future employment statuses; the DSS uses the sum of two of the outputs, corresponding to the probability of being employed within 6 months, either permanently or on fixed-term/part-time (Bennmarker et al., 2021).

The statistically estimated probability is combined with the jobseeker's current unemployment duration using threshold functions into three possible outcomes (Arbetsförmedlingen, 2020; see figure 1):

- Too near the job market – the jobseeker is deemed capable of finding a job with minor help, such as digital services
- Suitable for Prepare and Match
- Too far away from the job market – the jobseeker needs further investigation and other kinds of support

The thresholds between different outcomes are subject to political or administrative decisions related to e.g. available resources.

The system's accuracy, measured as the fraction of historical data points that are assigned an adequate decision (i.e. positive decision for jobseeker without job after 6 months, and vice versa), is 68%. Accuracy differs across sub-populations and decisions; the lowest accuracy is reported for negative decisions for jobseekers with disabilities (F1=17%) (Böhlmark, Lundström & Ornstein, 2021).

[2] https://www.esv.se/statsliggaren/regleringsbrev/?RBID=20264 (Accessed Jan 19, 2022)

[3] This can be contrasted with an earlier Swedish system, where caseworkers were instructed to consider the recommended decision carefully but to also use their professional judgement (Assadi & Lundin, 2018).

[4] The neural net has 64 inputs, two hidden layers and 14 outputs.

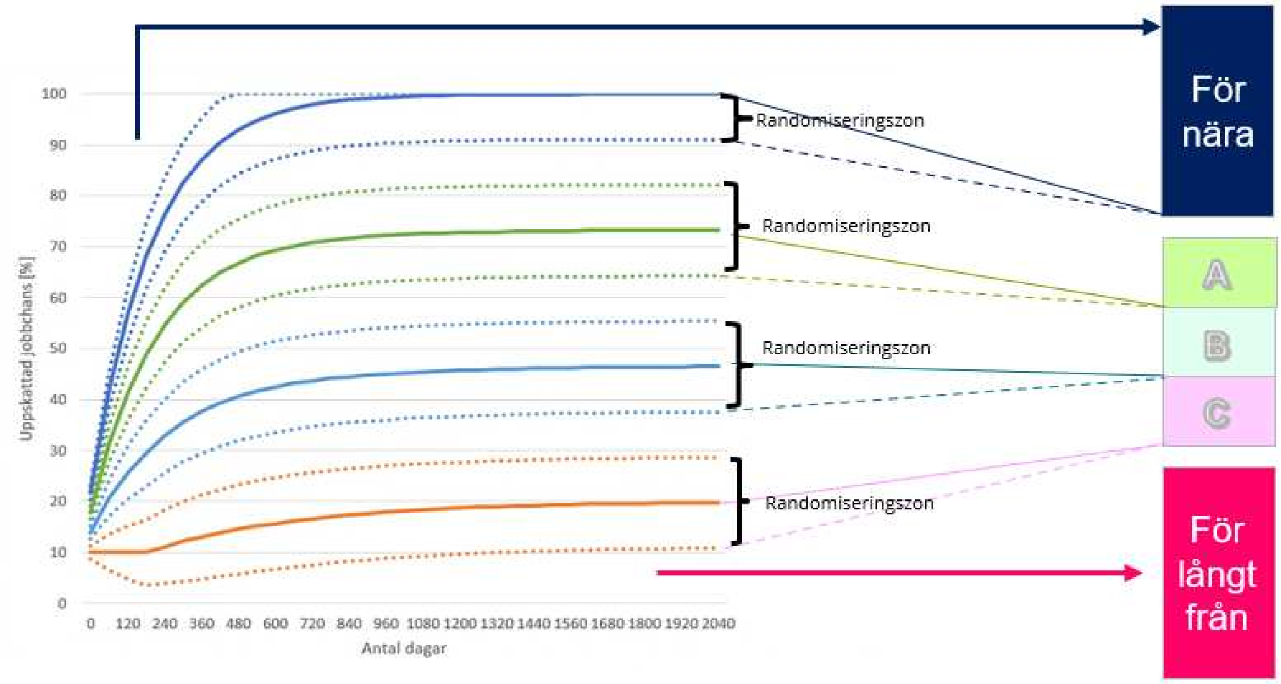

Figure 1. Thresholds and outcomes. Relationship between estimated probability of finding a job ("uppskattad jobbchans"), number of days of current unemployment ("antal dagar"), and outcome (too near ("för nära") or too far away from ("för långt ifrån") the labour market, or positive decision). For example, if job chance is estimated at 50% and unemployment duration is 360 days, the jobseeker is recommended access to the employment initiative. (Levels A-C affect the amount of compensation that providers receive.) Note that this illustration is not presented to jobseekers as part of any explanation. Reprinted with permission from Arbetsförmedlingen.

3.3 Explanations

Decisions about access to the employment initiative are communicated to the jobseeker in a meeting with a caseworker. Towards caseworkers, the recommended decision is shown in the case management system and is accompanied by a ranking of the 10 most important factors. The decision is also sent as a letter to the jobseeker and presented to the jobseeker when logged in at the agency's website. Towards jobseekers, only the top 4 most important factors are listed.

A suggested phrasing of the decision is automatically generated by the case management system (Arbetsförmedlingen, 2020). Below is an example of a positive decision (my translation):

> By comparing your information with statistics we have tried to assess how near you are the job market. Our assessment is that you will get the best help from a supervisor at one of the providers within the initiative Prepare and Match. In your case it was primarily the following factors that contributed to the assessment: Your unemployment duration, Your unemployment history, Your city of residence and Working time.

Unemployment duration is always presented as the most important factor. The rest of the factors are ranked using a method called LIME (Ribeiro, Singh & Guestrin, 2016). Given an input (i.e. data for a jobseeker at a particular point in time), LIME creates a simplified model by systematically investigating outcomes for various modifications of the input. For example, if age is assigned a high rank by LIME in a given case, it means that in situations similar to the case at hand, a different age tends to cause a different outcome.

Neither the estimated probabilities of different outcomes, the system's accuracy, or the thresholds and their influence on decisions are communicated to jobseekers or caseworkers. Implications of omitting such information will be discussed in section 4.

4 EXPLAINABILITY ANALYSIS

Previous studies of the DSS have shown that agency officials find it difficult to understand the basis for specific decisions, sometimes referring to the system as a black box (Bennmarker et al., 2021; Carlsson, forthcoming). The analysis below may illuminate why this is the case, although it takes the perspective of affected jobseekers rather than caseworkers. Following Scott et al. (2022), the analysis focuses on jobseekers' interests in *intelligibility* (outputs from system should be understandable) and *empowerment* (system should empower the jobseeker e.g. by providing actionable information).

4.1 Opaque internal logic

The statistical model is a neural network which, due to its non-linear processing and complex interactions between variables, is fairly opaque. Consequently, it is difficult even for AI experts with full access to the model to understand how the model reaches its judgements. This circumstance underpins many of the other issues raised below.

4.2 Unreliable explanation method

A common approach for explaining predictions by opaque models is to create a simpler, interpretable model that approximates the opaque model on a case-by-case basis, and then get explanations from the "surrogate" model instead. The agency uses one of the most popular techniques of this kind, called LIME (Ribeiro, Singh & Guestrin, 2016), to rank importance of factors.

While LIME and similar methods can give some insight into how an opaque model operates, the methods have been shown to be unstable: different explanations can be generated for the same prediction. Furthermore, since LIME and similar methods are approximate, explanations are not always faithful with respect to the outcomes that they are supposed to explain (Amparore, Perotti & Bajardi, 2021). In other words, the potential intelligibility afforded by approximate explanations comes at the cost of unreliability.

4.3 Misleading importance attribution for unemployment duration

In addition to unreliability issues associated with the explanation method as such, the special treatment of unemployment duration raises additional concerns. Towards jobseekers, the list of factors is presented as case-specific ("*In your case* it was primarily the following factors..."). However, this is misleading in the sense that current duration of unemployment is programmed to always appear first in the list. Furthermore, the special treatment leads to potentially inaccurate explanations, since the importance of unemployment duration may vary from case to case. As illustrated by figure 1, the effect of unemployment duration on decisions diminishes as duration increases. For example, we can consider a jobseeker that is deemed too far away from the job market, has been unemployed for 2000 days and is near the decision threshold (i.e. the estimated probability of finding a job is slightly below 20%). In such a situation, a positive decision would have required a much shorter unemployment duration, or just a slight increase in estimated probability of finding a job (i.e. a potentially small change among other factors). In other words, there may exist cases where unemployment duration is less important than other factors.

4.4 Limited usefulness

Beyond issues regarding unreliability, previous work has shown that outputs from LIME and similar explanation methods can be difficult to interpret (Dieber & Kirrane, 2020). If city of residence is presented with a higher rank than working time for a negative outcome, what does this mean? Technically, the answer is that changing city of residence is more likely to lead to a positive outcome than changing working time. However, this information has limited value. For example, it does not explain *how* city of residence or working time would need to change in order to yield a positive decision.

Generally speaking, factor rankings do not enable the kind of counterfactual or contrastive reasoning that are common in human explanations. Research in linguistics and psychology has shown that humans tend to explain events in terms of conditions that would cause another event to occur (Miller, 2019). In the context of the current case, a counterfactual explanation for a negative decision could be expressed as: "If you would seek a full-time rather than part-time employment, your chances of finding a job would likely increase and you would be considered near enough the job market to get help within the initiative Prepare and Match". As argued by Wachter, Mittelstadt & Russell (2017), counterfactual explanations not only convey why or how a particular decision was reached, but also provide grounds to contest a decision and guidance on how to receive a different (e.g. more desired) outcome in the future. A similar recommendation is made by European Parliamentary Research Service in relation to automated decision-making, arguing that "data subjects who did not obtain the decision they hoped for should be provided with the specific information that most matters to them, namely, with the information on *what values for their features* determined in their case an unfavourable outcome" (Sartor & Lagioia, 2020, emphasis mine).

Since counterfactual explanations depend on notions of actionability that may differ between subjects (Rudin, 2019) – for example, switching from part-time to full-time work may be more feasible for some jobseekers than others – counterfactual explanations may require some kind of interaction between system and jobseeker (see section 4.6).

4.5 Choice of model

Several of the issues discussed above boil down to the opacity of the statistical model. Two international comparisons can illustrate how a more transparent model could potentially mitigate these issues. The Danish PES has used a decision tree with only five variables and very few interactions between variables. For example, if a jobseeker is unconfident about finding a job, the model predicts a 83% risk of future unemployment, regardless of other factors; if the jobseeker is more optimistic, the model uses three additional factors (age, previous employment rate and migration status) to categorise risk of unemployment into three different probabilities.[5] The Polish PES has used an algorithm with 24 questions scored from 0 (highest employability) to 8. Depending on the total score, the

[5] https://star.dk/media/12514/2020_01_31_beskrivelse_-_profilafklaringsvaerktoej_til_dagpengemodtagere.pdf (Accessed Feb 17, 2023)

jobseeker is categorised into one of three profiles (Sztandar-Sztanderska & Zielenska, 2020). In both the Danish and Polish case, the simplicity of the model eliminates the need of an additional explanation method; the models more or less "explain themselves". For example, if a Danish jobseeker wants to know why the model makes a particular prediction, a caseworker can show the decision tree in its entirety and highlight the path at hand. Seeing the entire decision tree also enables counterfactual reasoning, since it is easy to see how an alternative path leads to a different outcome. Similarly, if a Polish jobseeker wants to know how to increase his/her job chances (according to the algorithm), this information is directly contained in the scoring of individual questions. Note however that this requires that the scoring criteria are disclosed, which has not been the case with the Polish system (Niklas et al., 2015).

Is a simple and interpretable model, such as a small decision tree or a scoring algorithm, as accurate as a more opaque neural network? Comparing accuracy across countries is difficult, since the data varies between the countries. However, the Swedish PES has experimented with two models that are much simpler and explainable than the deployed one, and whose accuracy can be compared to the deployed model using the same data. The simplest model (a linear regressor), has an accuracy of 66%, comparable with the 68% for the deployed model (Ornstein & Thunström, 2021); a slightly more sophisticated model (small decision tree + 6 linear regressors) has an accuracy of 74% (Helgesson & Ornstein, 2021), i.e. *better* than the deployed model. This suggests that a simpler model can fulfil the stated goals – consistency and accuracy – equally well, or even better, than an opaque model, without the negative consequences for explainability that an opaque model brings about. This finding also resonates with some previous work on the relationship between accuracy and interpretability (Rudin, 2019).

4.6 Interactivity

Philosophical, cognitive, and social studies of explanations tend to emphasise their social nature: explanations involve transfer of knowledge in an interaction between an explainer and an explainee (Miller, 2019). In line with this, some scholars emphasise the potential values of interactive explanations (Miller, 2019; Arya et al., 2019; Weld & Bansal, 2019, Simkute et al. 2021; Lakkaraju et al. 2022; Berman & Howes, 2022; Cheng et al., 2019). For example, interactivity can enable stakeholders to ask "what-if" questions for hypothetical circumstances, without any need for simplified approximations (Wachter et al., 2017). If a jobseeker is denied access to the employment initiative, the possibility to ask questions such as "What if I move to Stockholm?" or "What if I get a university degree?" can help the jobseeker to not only understand how the AI makes its judgements, but also to use this understanding to negotiate or contest a decision. To the extent that the AI has learned something relevant about employability, exploration of hypothetical circumstances also enables the AI to be used as a coach for getting advise on how to get nearer the job market (Scott et al., 2022).

Supporting hypothetical questions is technically trivial; users only need to be equipped with a graphical interface which allows exploring how modifying the input affects the output. Interactivity could in principle also enable more open-ended counterfactual questions such as "What would motivate a positive decision in my case?", where the feasibility of changes in circumstances can be addressed in a dialogue between the system and jobseeker (Berman et al., 2022).

4.7 Accuracy

As mentioned in section 3.3, the accuracy of the system is not communicated to jobseekers or caseworkers, despite the fact that accuracy is far from perfect and varies greatly across different subpopulations and decisions. This makes it difficult for jobseekers to assign adequate degrees of trust in the AI. For example, to the extent that the AI can be used for getting actionable advise, knowledge about accuracy enables jobseekers to assess the reliability of the advise. Accuracy information also helps jobseekers to assess how appealable their case is; in situations where the AI is less accurate, there might be more room for negotiation.

To mitigate this, stakeholders could be provided with performance indicators for the relevant subpopulation and decision. For example, if a jobseeker with disabilities is rejected access to the initiative, the system could provide a reservation about its high uncertainty.

4.8 Thresholds and confidence estimates

As described in section 3.2, outcomes are partly governed by thresholds that are continuously adjusted by the agency. For example, if the agency lowers the threshold for positive decisions, some jobseekers may obtain a positive decision as a direct consequence of the changed threshold. However, the thresholds are mentioned neither in explanations for specific decisions or in general information to the public on the agency's web site. Arguably, concealing some of the factors that underpin decisions impedes jobseekers' ability to understand the basis for the decisions.

Furthermore, the probabilities of future employment statuses estimated by the statistical model are not communicated to stakeholders. As with accuracy information (see section 4.7), this makes it difficult for jobseekers to assess how much individual assessments can be trusted. Arguably, jobseekers have an interest in knowing if their decision is considered straightforward and univocal, or if it is a borderline case with high uncertainty. For example, if the model predicts a job chance of 5% for jobseeker A and 20% for jobseeker B, then both jobseekers are deemed too far away from the job market (assuming that they are both long-term unemployed). Nevertheless, jobseeker B is very near the threshold for a positive decision, and should therefore be in a more negotiable situation.

In this regard, explainability could potentially be enhanced by showing a simplified variant of figure 1, where the relevant region of the decision landscape has been zoomed in and/or highlighted. Additionally, the probability of making the correct decision given the current thresholds can be calculated and presented. For example, the confidence value would be near 50% for person B (indicating a very low confidence of recommending the right decision), while it would be higher for person A.

5 CONCLUSIONS

This case study of an AI-based decision-support system deployed by the Swedish Public Employment Service has shown that its justifications of decisions lack important information and are unreliable, potentially misleading and difficult to interpret. These weaknesses in explainability may affect jobseekers by influencing the caseworkers' decision-making; if caseworkers had access to more intelligible and reliable explanations, this might have affected their trust in the AI in either direction from case to case, and thereby also the final decisions. First and foremost, however, the study has highlighted how the weaknesses impede jobseekers' ability to understand, negotiate and contest dispreferred decisions, and to get advise on how to increase their employment chances.

The good news is that many of the highlighted issues could be mitigated by replacing the current opaque statistical model with a simpler, more interpretable one; this would address jobseekers' interests and needs without necessarily impairing other desiderata. Increasing the degree of interactivity could also serve jobseekers' needs, potentially without replacing the statistical model.

It is important to note that the jobseeker perspective adopted in the present study is based on insights from previous research involving jobseekers in somewhat different contexts (Scott et al., 2022). In future work, it would be useful to empirically study the extent to which jobseekers find provided explanations intelligible and useful, e.g. using questionnaires and interviews. (Such studies could potentially also involve alternative, e.g. more interactive, forms of explanations.) It would also be interesting to collect and analyse caseworker-jobseeker conversations and study their strategies in relation to the AI.

Finally, it should be stressed that explainability is only one of many aspects to consider when assessing a decision-support system (other aspects include e.g. fairness). Nonetheless, this study may contribute to a better understanding of how choice of statistical model and design of explanations can impact the value and usefulness of an AI-based decision-support system from the perspective of those that are directly affected by the decisions; these insights may be relevant in other domains as well.

ACKNOWLEDGEMENTS

This work was supported by the Swedish Research Council (VR) grant 2014-39 for the establishment of the Centre for Linguistic Theory and Studies in Probability (CLASP) at the University of Gothenburg.

REFERENCES

1. Allhutter, Doris, Florian Cech, Fabian Fischer, Gabriel Grill, and Astrid Mager. 2020. "Algorithmic profiling of job seekers in Austria: how austerity politics are made effective." *Frontiers in Big Data*, 5.
2. Amparore, Elvio, Alan Perotti, and Paolo Bajardi. 2021. "To trust or not to trust an explanation: using LEAF to evaluate local linear XAI methods." *PeerJ Computer Science* 7:e479.
3. Arbetsförmedlingen. 2020. *Arbetsförmedlingens handläggarstöd.* Dnr Af-2020/0016 7459.
4. Arya, Vijay, Rachel KE Bellamy, Pin-Yu Chen, Amit Dhurandhar, Michael Hind, Samuel C Hoffman, Stephanie Houde, Q Vera Liao, Ronny Luss, Aleksandra Mojsilović, et al. 2019. "One explanation does not fit all: A toolkit and taxonomy of AI explainability techniques." *arXiv preprint arXiv:1909.03012.*
5. Assadi, Anahita, and Martin Lundin. 2018. "Street-level bureaucrats, rule-following and tenure: How assessment tools are used at the front line of the public sector." *Public Administration* 96 (1): 154–170.
6. Bennmarker, Helge, Martin Lundin, Tove Mörtlund, Kristina Sibbmark, Martin Söderström, and Johan Vikström. 2021. *Krom – erfarenheter från en ny matchningstjänst med fristående leverantörer inom arbetsmarknadspolitiken.* Institutet för arbetsmarknads- och utbildningspolitisk utvärdering (IFAU), July.
7. Berman, Alexander, Ellen Breitholtz, Jean-Philippe Bernardy, and Christine Howes. 2022. "Explaining Predictions with Enthymematic Counterfactuals." In *Proceedings of the 1st Workshop on Bias, Ethical AI, Explainability and the role of Logic and Logic Programming, BEWARE-22*, 95–100.

8. Berman, Alexander, and Christine Howes. 2022. “"Apparently acousticness is positively correlated with neuroticism". Conversational explanations of model predictions.” In *Proceedings of SEMDIAL 2022 (DubDial)*.
9. Böhlmark, Anders, Tom Lundström, and Petra Ornstein. 2021. *Träffsäkerhet och likabehandling vid automatiserade anvisningar inom Rusta och matcha. En kvalitetsgranskning.* Arbetsförmedlingen analys.
10. Carlsson, Vanja. Forthcoming.
11. Carney, Terry. 2020. “Artificial intelligence in welfare: Striking the vulnerability balance?” *Monash University Law Review* 46 (2): 23–51.
12. Cheng, Hao-Fei, Ruotong Wang, Zheng Zhang, Fiona O’Connell, Terrance Gray, F Maxwell Harper, and Haiyi Zhu. 2019. “Explaining decision-making algorithms through UI: Strategies to help non-expert stakeholders.” In *Proceedings of the 2019 CHI conference on human factors in computing systems*, 1–12.
13. Dastile, Xolani, Turgay Celik, and Moshe Potsane. 2020. “Statistical and machine learning models in credit scoring: A systematic literature survey.” *Applied Soft Computing* 91:106263.
14. Desiere, Sam, Kristine Langenbucher, and Ludo Struyven. 2019. “Statistical profiling in public employment services,” *OECD Social, Employment and Migration Working Papers*, no. 224, https://doi.org/10.1787/b5e5f16e-en.
15. Desiere, Sam, and Ludo Struyven. 2021. “Using artificial intelligence to classify jobseekers: the accuracy-equity trade-off.” *Journal of Social Policy* 50 (2): 367–385.
16. Dieber, Jürgen, and Sabrina Kirrane. 2020. *Why model why? Assessing the strengths and limitations of LIME.* https://doi.org/10.48550/ARXIV.2012.00093.
17. Hansson, Ewa, Gioia Luigetti, Martin Waara, and Stefan Öster. 2022. *ESF-projekt Kundval rusta och matcha. Slutrapport.* Arbetsförmedlingen.
18. Helgesson, Petter, and Petra Ornstein. 2021. *Vad avgör träffsäkerheten i bedömningar av arbetssökandes stödbehov? En undersökning av förutsättningarna för statistiska bedömningar av avstånd till arbetsmarknaden, med fokus på betydelsen av inskrivningstid.* Arbetsförmedlingen analys.
19. Lakkaraju, Himabindu, Dylan Slack, Yuxin Chen, Chenhao Tan, and Sameer Singh. 2022. “Rethinking Explainability as a Dialogue: A Practitioner’s Perspective.” *arXiv preprint arXiv:2202.01875.*
20. Miller, Tim. 2019. “Explanation in artificial intelligence: Insights from the social sciences.” *Artificial intelligence* 267:1–38.
21. Niklas, Jędrzej, Karolina Sztandar-Sztanderska, Katarzyna Szymielewicz, A Baczko-Dombi, and A Walkowiak. 2015. *Profiling the unemployed in Poland: social and political implications of algorithmic decision making.* Fundacja Panoptykon.
22. Ornstein, Petra, and Hanna Thunström. 2021. *Träffsäkerhet i bedömningen av arbetssökande. En jämförelse av arbetsförmedlare och ett statistiskt bedömningsverktyg.* Arbetsförmedlingen analys.
23. Qayyum, Adnan, Junaid Qadir, Muhammad Bilal, and Ala Al-Fuqaha. 2021. “Secure and Robust Machine Learning for Healthcare: A Survey.” *IEEE Reviews in Biomedical Engineering* 14:156–180. https://doi.org/10.1109/RBME.2020
24. Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin. 2016. “"Why Should I Trust You?": Explaining the Predictions of Any Classifier.” In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1135–1144. KDD ’16. San Francisco, California, USA: Association for Computing Machinery. https://doi.org/10.1145/2939672.2939778
25. Rudin, Cynthia. 2019. “Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead.” *Nature Machine Intelligence* 1 (5): 206–215.
26. Sartor, Giovanni, and Francesca Lagioia. 2020. *The impact of the General Data Protection Regulation (GDPR) on artificial intelligence.* European Parliamentary Research Service.
27. Saxena, Devansh, Karla Badillo-Urquiola, Pamela J. Wisniewski, and Shion Guha. 2020. “A Human-Centered Review of Algorithms Used within the U.S. Child Welfare System.” In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 1–15. CHI ’20. Honolulu, HI, USA: Association for Computing Machinery. https://doi.org/10.1145/3313831.3376229

28. Scott, Kristen M, Sonja Mei Wang, Milagros Miceli, Pieter Delobelle, Karolina Sztandar-Sztanderska, and Bettina Berendt. 2022. "Algorithmic Tools in Public Employment Services: Towards a Jobseeker-Centric Perspective." In *2022 ACM Conference on Fairness, Accountability, and Transparency*, 2138–2148.

29. Simkute, Auste, Ewa Luger, Bronwyn Jones, Michael Evans, and Rhianne Jones. 2021. *"Explainability for experts: A design framework for making algorithms supporting expert decisions more explainable."* Journal of Responsible Technology 7-8:100017. https://doi.org/10.1016/j.jrt.2021.100017

30. Sztandar-Sztanderska, Karolina, and Marianna Zielenska. 2020. "What Makes an Ideal Unemployed Person? Values and Norms Encapsulated in a Computerized Profiling Tool." *Social Work & Society* 18(1) (May).

31. Sztandar-Sztanderska, Karolina, and Marianna Zielenska. 2022. "When a Human Says "No" to a Computer: Frontline Oversight of the Profiling Algorithm in Public Employment Services in Poland." *Sozialer Fortschritt* 71 (6-7): 465–487.

32. Vaassen, Bram. 2022. "AI, Opacity, and Personal Autonomy." *Philosophy & Technology* 35 (4): 88.

33. Wachter, Sandra, Brent Mittelstadt, and Chris Russell. 2017. "Counterfactual explanations without opening the black box: Automated decisions and the GDPR." *Harv. JL & Tech.* 31:841.

34. Weld, Daniel S, and Gagan Bansal. 2019. "The challenge of crafting intelligible intelligence." *Communications of the ACM* 62 (6): 70–79.

35. Zejnilovic, Leid, Susana Lavado, Carlos Soares, Íñigo Martınez De Rituerto De Troya, Andrew Bell, and Rayid Ghani. 2021. "Machine Learning Informed Decision-Making with Interpreted Model's Outputs: A Field Intervention." In *Academy of Management Proceedings*, 2021:15424. 1. Academy of Management Briarcliff Manor, NY 10510.