Trustworthy AI in the public sector: An empirical analysis of a Swedish labor market decision-support system

Technology in Society 76 (2024) 102471

Contents lists available at ScienceDirect

Technology in Society

journal homepage: www.elsevier.com/locate/techsoc

Trustworthy AI in the public sector: An empirical analysis of a Swedish labor market decision-support system

Alexander Berman [a,*], Karl de Fine Licht [b], Vanja Carlsson [c]

[a] Centre for Linguistic Theory and Studies in Probability (CLASP), Department of Philosophy, Linguistics and Theory of Science, University of Gothenburg, Box 200, 405 30, Gothenburg, Sweden
[b] Division of Science, Technology and Society, Chalmers University of Technology, Sweden
[c] School of Public Administration, University of Gothenburg, Sweden

ARTICLE INFO

Keywords:
Decision-support systems
Trustworthy AI
Artificial Intelligence
Public employment services
Public sector

ABSTRACT

This paper investigates the deployment of Artificial Intelligence (AI) in the Swedish Public Employment Service (PES), focusing on the concept of trustworthy AI in public decision-making. Despite Sweden's advanced digitalization efforts and the widespread application of AI in the public sector, our study reveals significant gaps between theoretical ambitions and practical outcomes, particularly in the context of AI's trustworthiness. We employ a robust theoretical framework comprising Institutional Theory, the Resource-Based View (RBV), and Ambidexterity Theory, to analyze the challenges and discrepancies in AI implementation within PES.

Our analysis shows that while AI promises enhanced decision-making efficiency, the reality is marred by issues of transparency, interpretability, and stakeholder engagement. The opacity of the neural network used by the agency to assess jobseekers' need for support and the lack of comprehensive technical understanding among PES management contribute to the challenges in achieving transparent and interpretable AI systems. Economic pressures for efficiency often overshadow the need for ethical considerations and stakeholder involvement, leading to decisions that may not be in the best interest of jobseekers.

We propose recommendations for enhancing AI's trustworthiness in public services, emphasizing the importance of stakeholder engagement, particularly involving jobseekers in the decision-making process. Our study advocates for a more nuanced balance between the use of advanced AI technologies and the leveraging of internal resources such as skilled personnel and organizational knowledge. We also highlight the need for improved AI literacy among both management and personnel to effectively navigate AI's integration into public decision-making processes.

Our findings contribute to the ongoing debate on trustworthy AI, offering a detailed case study that bridges the gap between theoretical exploration and practical application. By scrutinizing the AI implementation in the Swedish PES, we provide valuable insights and guidelines for other public sector organizations grappling with the integration of AI into their decision-making processes.

1. Introduction

In recent years, we have witnessed a proliferation of AI systems in public sectors across the globe. The driving force behind this phenomenon has been the aspiration to improve decision-making processes, making them more uniform, accurate, and efficient, and therefore enhancing the productivity and reliability of public services (see e.g., Refs. [1–4]). However, this rapid integration of AI in our societal frameworks has also necessitated a stringent focus on the aspect of "trustworthiness", emphasizing that a trustworthy AI system should ideally be transparent, explicable, lawful, ethical, and robust in its functionalities (see e.g. Ref. [3]). This belief was concretized by the EU's High-Level Expert Group on AI when they published "Ethics Guidelines for Trustworthy Artificial Intelligence" [5], a document that has since become a cornerstone for studies examining the practicalities of trustworthiness in AI [6–8].

Yet, a significant gap remains. Despite extensive discourse around trustworthy AI, we lack a comprehensive, in-depth empirical analysis of real-life scenarios in which both the technological intricacies of AI systems and their social implications are examined concurrently. It is not

* Corresponding author.
E-mail addresses: alexander.berman@gu.se (A. Berman), karl.definelicht@chalmers.se (K. de Fine Licht), vanja.carlsson@spa.gu.se (V. Carlsson).

https://doi.org/10.1016/j.techsoc.2024.102471
Received 18 July 2023; Received in revised form 24 January 2024; Accepted 24 January 2024
Available online 26 January 2024

only necessary to identify these shortcomings, but we also need to provide practical solutions that do not demand a complete cessation of AI usage or entirely new technical developments. While strides have been made in elucidating how the use of AI can be optimized to embody our societal principles and values, both broadly and specifically within the public sector [9,10], and robust studies examining the societal impact of AI exist (see e.g. Ref. [11]), alongside overviews of trustworthy AI where problems and potential solutions are discussed in the abstract [3] and in terms of practical methodology [12], there remains an urgent need for tangible case studies where these things are brought about together and further examined in concrete terms (see e.g. Refs. [4, 13]). Such work would render the concept of trustworthy AI more operational and applicable to specific governance situations.

In this paper, we bridge this gap. We delve into the case of AI usage by the Swedish Public Employment Service (PES) and apply some of the most uncontroversial principles of "trustworthy AI". Our examination covers a spectrum of aspects, such as whether the AI is explainable and interpretable, and to what extent it contributes to fair and equal treatment. Our exploration reveals an array of challenges. For example, the AI, a neural network, is inscrutable (a 'black box'), its explanations inadequate, and it presents difficulties in its role within the decision-making process, especially the ability to contest decisions. In response to these findings, we propose potential solutions to enhance the decision-making process and bolster the trustworthiness of AI. These include increased participation of jobseekers, an expansion of professional discretion, improvement in performance indicators, and a simplification of the conceptualization of the decision-making logic.

The structure of our paper is as follows: In Section 2, we present our main theories, define key concepts, and sketch a framework for assessing an AI system's trustworthiness within the context of public decision-making. The research methodology and empirical material is presented in Section 3. In Section 4, we delve into a case of AI-assisted decision-making in the Swedish Public Employment Service and apply our framework to assess the system's trustworthiness and discuss ways to improve it. The results are then analyzed theoretically in Section 5. In Section 6, we discuss theoretical implications, make recommendations, and discuss relation with earlier work and limitations of the study. Finally, Section 7 concludes with a summary of our findings and a reflection on the future landscape of trustworthy AI in public decision-making.

2. Trustworthy AI

In this section we are first going to discuss the main theories that can help us explain why and how we might achieve trustworthy AI and then we will discuss how trustworthy AI should be defined. The theoretical framework in the next section will primarily be used in the discussion section in addition to informing the collection of our empirical material.

2.1. Theoretical framework

When analyzing the use of Artificial Intelligence (AI) in public decision-making, it is crucial to frame our discussion within a robust theoretical context. To this end, we draw upon three well-established theoretical frameworks, each offering unique insights into the organizational and technological dynamics at play successfully used by Di Vaio and colleagues [4] in a similar context.

Firstly, Institutional Theory [14] sheds light on how societal norms, rules, and expectations influence organizational behavior and decision-making processes in the public sector. This perspective is particularly relevant for understanding the Swedish Public Employment Service's adoption and integration of new technologies like AI and Big Data (BD), influenced by both external pressures and internal dynamics.

Secondly, the Resource-Based View (RBV) [15] emphasizes the significance of leveraging internal resources, including technological infrastructure, skilled personnel, and organizational knowledge. This view is crucial in understanding how these internal assets are utilized to harness the capabilities of AI, BD, and Data Intelligence and Analytics (DI&A) for enhancing public sector decision-making processes.

Lastly, Ambidexterity Theory [16] explores the balancing act between exploiting existing resources and exploring new technological opportunities. This theory is key to comprehending how the Swedish Public Employment Service maintains and how it could produce efficient operations while integrating emerging technologies like DI&A, AI, BD, and augmented decision-making, or Human-Artificial Intelligence (HAI), where humans are using AI in analyzing data and taking decisions.

Building on these theoretical frameworks, we delve into the core technological concepts of DI&A, AI, BD, and HAI. In the context of the Swedish Public Employment Service, these technologies are not standalone tools, but part of a larger system intertwined with organizational practices and policies. DI&A plays a pivotal role in processing and analyzing vast amounts of data, AI is hoped to enhance decision-making accuracy and efficiency, and BD represents the extensive data landscape that feeds into these processes.

This interplay between the theoretical frameworks and technological concepts forms the foundations of our exploration into the Swedish Public Employment Service's application of AI. We critically examine how the technological concepts and the theoretical frameworks collectively contribute to our understanding of what the challenges are regarding the trustworthiness of AI systems in public decision-making and how these can be met. Our analysis covers aspects such as the explainability and interpretability of AI decisions, the system's alignment with legal and ethical standards, and the integration of AI within the broader organizational context.

2.2. Principles for trustworthy AI

Numerous frameworks for trustworthy AI have been proposed, each suggesting different criteria (for a recent overview, see Ref. [3]). Although our discussion may not encompass all possible perspectives, we believe that the principles or criteria presented here are widely accepted, and therefore serve as a good starting point for any case study on trustworthy AI. The principles outlined in this paper have previously been proposed in the literature, such as the report on trustworthy AI sponsored by the EU Commission [5] as well as reviews of ethics guidelines [17] and trustworthy AI [3].

Our analysis will focus on six key principles. The first is performance, which in the studied case can be evaluated at various levels including system-wide (general predictions), sub-population specific (e.g., jobseekers with and without disabilities), and outcomes (e.g., positive and negative decisions). These levels are crucial from a public policy perspective, so we will examine all of them here. It should also be noted that when AI assists human decisions, the performance of human decisions may differ from the performance of the AI alone as well as from human decisions without AI assistance. For example, in the studied case, caseworkers are instructed to primarily follow the automated recommendation but may override it in individual cases. Therefore, the performance of the caseworker when assisted by the system – a concept we term "augmented performance" – may differ from the performance of the system alone [18]. As such, it becomes necessary to evaluate not just the performance of the AI system alone, but also how it affects the performance of the human that uses the AI (in this case the caseworker). Such effects can be measured e.g. via reliance patterns or effects on accuracy [19].

The second principle is that of *calibration* [20,21]. For stakeholders, it is crucial to know how confident an AI system is about a particular decision or judgment. Calibration refers to a system's ability to correctly estimate this confidence. A well-calibrated system will correctly estimate its confidence most of the time. Conversely, a poorly calibrated system may either over- or underestimate its confidence. In principle, a system may perform satisfactorily (e.g., better than some baseline) but

still be ill-calibrated. For example, it may ascribe a high probability even when it is wrong, or a low probability even when it is correct.

All else being equal, a well-calibrated system that communicates its confidence is more reliable, as it allows users to apply appropriate trust on predictions on a case-by-case basis. For example, if an AI is very uncertain about a specific prediction, it may be adequate for the human decision-maker to not rely on the AI, even if the AI generally performs well. Furthermore, from the perspective of subjects of decisions, understanding whether a decision is considered straightforward and univocal, or a borderline case with high uncertainty, can help subjects assess if there is room for negotiation and appeal. Thus, both having a well-calibrated system as well as communicating confidence levels are key to a trustworthy AI system in public decision-making.

The third and fourth principles encompass *interpretability*, *explainability, intelligibility,* and *availability*. Interpretability refers to the ability of humans to in principle understand the logic behind an AI's decisions and outcomes [22,23]. This quality varies between different types of AI. Classical AI, for instance, was entirely based on rules that formalized domain knowledge in algorithms understandable to anyone. Modern AI, however, often relies on machine learning, meaning that algorithms guide the AI's learning from examples rather than transparently encoding domain knowledge. The resulting logic can vary in interpretability, with some models (e.g., small decision trees) generating comprehensible rules, while others, like neural networks, may be seen as inscrutable "black boxes" [22]. Interpretability can be evaluated experimentally with end users or on the basis of formal criteria [24].

Even if an AI system lacks interpretability, it may still be possible to approximate its internal logic to produce explanations for specific judgments or decisions. This property is called "explainability" (see e.g. Ref. [25]). For instance, a complex neural network can be simplified as a linear model for specific decisions. Conceptually, one can say that the opaque AI is explained by an additional AI that tries to simplify the logic of the opaque one. However, since such explanation methods approximate the actual decision-making logic, they might not always be faithful with respect to the outcomes that they are supposed to explain [26]. In general, we have weaker reasons to trust an AI system that uses approximate explanations over one with interpretable decision-making logic, especially when the approximations have low fidelity. A system's degree of explainability can be evaluated experimentally with end users [24]. Explanation methods can also be evaluated in terms of computational metrics such as local concordance and fidelity [26].

Intelligibility, closely related to interpretability and explainability, concerns whether the decision-making logic is communicated and presented in a way that is comprehensible in practice to those who have an interest in understanding the decisions. Thus, it is crucial not only that the AI system is interpretable or explainable but also that the design of AI-based judgments and decisions is informed by knowledge about how humans process and understand explanations [27]. Intelligibility can be evaluated experimentally with end users [24] or on the basis of theoretical criteria [27]. *Availability* has to do with making relevant information available for stakeholders and can be evaluated qualitatively by assessing to what extent relevant information is provided to stakeholders.

The fifth principle underscores the necessity for *fair and equal treatment* of subjects by AI systems. Equal treatment requires that AI systems treat cases that are relevantly similar in a similar manner [28], suggesting that the system should not incorporate excessive randomness. If it does, it might reflect the unpredictable "noisy" behaviors of humans (see e.g. Ref. [29]). Furthermore, potential biases need to be examined, particularly regarding sex, transgender identity or expression, ethnicity, religion or other belief, disability, sexual orientation, and age. Discrimination based on these grounds is unlawful (2007/08:95). However, exactly how to define and measure fairness and equality in the context of AI remains an open question [30]. Furthermore, previous work has shown that central aspects of fairness and equality can be difficult or impossible to achieve at once [31]. For these reasons, evaluation metrics need to be chosen on a case-by-case basis.

The sixth principle covers *legality, negotiation, and appeal*. The use and construction of AI should adhere to the law. Many aspects discussed previously are closely related to legality. For example, the EU's General Data Protection Regulation (GDPR) stipulates a "right to explanation" for individuals subjected to automated decisions, and the proposed EU AI Act encompasses fairness/bias, transparency, and interpretability.[1] According to the Swedish Administrative Procedure Act (2017:900), a decision expected to significantly affect an individual's situation should include a clear statement of reasons. Laws and regulations often encompass notions of equal treatment, such as the Swedish Act Concerning Equality between Men and Women (1991:433). More generally, the ability to predict the outcome of actions is central to many definitions of the rule of law and legal certainty [32]. This concept of predictability is closely tied to the transparency of decision-making logic, intelligibility, negotiability, and appealability in our analysis.

Public decision-making should also encompass the capacity to appeal unfavorable decisions. Aspects discussed earlier, such as the AI's accuracy, confidence, transparency, and intelligibility of decision-making logic, can facilitate this process. These features enable subjects to comprehend the foundations of decisions and, if warranted, contest a decision, or understand conditions that could lead to a more favorable outcome. Perhaps most importantly, explanations for decisions should make it possible to understand the conditions that would motivate another (e.g., a more favorable) decision [33,34].

Finally, the last principle concerns accountability and human oversight, by which we mean that the system should be set up in a way that makes it possible for human decision-makers to be accountable for the decisions that they make with the help of AI. For example, if caseworkers are formally responsible for decisions, then they need to be able to somehow oversee the AI and make a final decision, taking into account the output from the AI but also other considerations that may be relevant.

Notably, we do not explicitly discuss principles such as beneficence, non-maleficence, privacy, autonomy (see e.g. Ref. [1]), or technical robustness. Nevertheless, we contend that these principles are implicitly addressed or deliberately excluded due to our desire to maintain depth and focus within the discussion at hand. The primary benefit is associated with enhanced efficiency, accuracy, and equitable treatment achieved by minimizing variability in decision-making, while non-maleficence aligns with avoiding harm. Autonomy is preserved through elements such as explainability, interpretability, intelligibility, negotiation, and appeal, which empower the subjects' agency within the decision-making process. Privacy is ensured by adherence to legality, while technical robustness is omitted as it is challenging to address within the context of this discussion (cf. [3]).

Finally, when assessing the trustworthiness of an AI system, it is important to compare decision-making scenarios without the use of AI. When trustworthiness is greater for AI-assisted decision-making, we have reasons to trust it and vice versa. Some might argue that an AI system only needs to be sufficiently reliable, not fully reliable, for us to rely on it. However, to determine a precise sufficiency threshold, it is often necessary to make comparisons, typically leading to a comparative state. If stakeholders have good reasons to trust an AI system to improve the outcomes and procedures regarding the principles listed above, or if some stay the same while others are improved, we consider it an instance of what this paper terms "trustworthy AI".

Considering all that has been said until this point, the questions that should be asked to determine whether we have reasons to believe that an AI system, or the combination of human decision-makers and AI, is

[1] European Commission, Proposal for a Regulation of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) and amending certain Union legislative acts (COM (2021) 206 final).

trustworthy can be found in Table 1.

3. Methods and materials

A case-study research design [35] was applied. Sweden is a fruitful case to explore as the Swedish government is seeking to become "the best in the world at seizing the opportunities of digitalization" [36], our translation] and digital technologies are currently implemented in several different areas in the public sector. Sweden is also one of the countries with the highest quality of government in the world, with a long tradition of transparency in public decision-making, and in relative terms have well-resourced public agencies [37]. In Sweden, the PES is an agency that have made the most progress in applying AI systems.

The empirical material was analyzed using the questions formulated in Table 1. The analysis partly draws on previous studies [38,39] and is based on a comprehensive amount of empirical material including interviews, internal working materials, and information received from the Swedish Public Employment Service (PES) via email (between November 2021 and December 2023). We also reference public sources, policy documents, and published reports. The text material was mainly written and published by state actors and agencies, such as the PES, the Institute for the Evaluation of Labor Market and Education Policy (IFAU) and the Swedish government's official reports (SOU). The policy documents and reports are accessible online, and, in the analysis we explicitly refer to several of them. We chose these documents because they represent official descriptions and evaluations of the system as well as its correspondence to public principles.

In 2021, we conducted 15 semi-structured interviews with managers, qualified strategists, and officials at the Employment Service across both national and regional levels in Sweden. Each interview lasted between 45 and 90 min, with 13 conducted via Zoom and two over telephone. The interview questions dealt with the background and process of the introduction of the system and with the qualitative difference between the system and the previous decision-making procedure. The interview questions also dealt with the correspondence between the system and public principles such as accountability, efficiency, and legal certainty.

Table 1
Criteria for evaluating trustworthy AI.

No.	Criteria	Evaluation Questions
1	**Performance**	a. How accurately does the AI make judgments or decisions on all levels? b. Do human decision-makers make more accurate decisions with the help of the AI system? c. Is the system's performance communicated to stakeholders?
2	**Calibration**	a. Are confidence estimates communicated to stakeholders? b. If so, are the confidence estimates well-calibrated?
3	**Interpretability and Explainability**	a. Can the decision-making logic in principle be understood by stakeholders? b. Are explanations faithful with respect to the actual decision-making logic?
4	**Intelligibility and Availability**	a. Is the decision-making logic made available to the various stakeholders? b. Are explanations comprehensible for stakeholders in practice?
5	**Equal and Fair Treatment**	a. Does the AI make decisions consistently? b. Are relevant aspects of fair treatment satisfied?
6	**Legality, Negotiation, and Appeal**	a. Does the use and functionality of the AI system comply with the law? b. To what extent does the AI system enable affected individuals to negotiate or appeal unfavorable decisions?
7	**Accountability and Human Oversight**	Are human decision-makers able to oversee the operation of the AI and make independent decisions on the basis of the system's output?

The respondents are in number equally representing managerial-strategical level and the level of caseworkers and the same questions were asked to all respondents. All the interviews were transcribed and translated from Swedish into English. Access to the AI model as such was also requested but not granted.[2]

To the best of our knowledge, jobseekers' experience of the system has not been studied. Hence, when we assess aspects such as intelligibility from the perspective of jobseekers, we use theoretical criteria. We also assume that jobseekers have less knowledge than agency officials about how the system works.

4. Case study

In this section, we will first describe the general case. Following this, we will examine the extent to which PES meets the criteria for trustworthy AI and, if it does not, explore how this can be achieved.

4.1. Case description

The PES is responsible for the Swedish public employment services and labor-market policy activities. Its overall objectives are to bring jobseekers and employers together in an effective manner and to contribute to long-term employment. In Swedish unemployment policy, the so-called "labor-market policy assessment" is an important instrument in moving the unemployed into work. According to the regulations that governed the PES at the time of the study, this assessment must be formulated with the participation of each jobseeker. It regulates the jobseeker's rights and obligations in relation to receiving support and determines the activities the jobseeker must undertake.

The employment initiative Prepare and Match was launched in 2020[3] and rolled out nationally in 2021, following a directive from the Swedish government in 2019 that a statistical assessment support tool should be developed as an integrated part of the operations of the PES in order to improve consistency and accuracy of labor-market related assessments, and thereby improve the efficiency of resource allocation.[4] Through Prepare and Match, enrolled jobseekers get support e.g., in the form of training or guidance from a chosen provider. Decisions about whether jobseekers should be subject to Prepare and Match are assisted by a decision-support tool, called BÄR.

Thus, the labor market policy assessments have shifted from being manually conducted by caseworkers to being primarily conducted with the help of a statistical profiling tool. When a jobseeker contacts the PES for support, a caseworker uses the statistical profiling tool, which provides a result regarding recommended activities in the form of one of the following outcomes:

1. Too near the job market – the jobseeker is deemed capable of finding a job on his/her own, with minor help, such as digital services.
2. Suitable for Prepare and Match.
3. Too far away from the job market – the jobseeker needs further investigation and other, more in-depth, kinds of support.

The decision-support system consists of a statistical model that estimates the jobseeker's probability of finding a job within 6 months and threshold functions that, given the jobseeker's current duration of

[2] In response to our request for access, the agency stated it has chosen not to disclose details about the model "in order to avoid manipulation of input and as protection against cyberattacks such as adversarial attacks" (our translation).

[3] A second version of Prepare and Match was introduced in April 2023. This paper focuses on the first version.

[4] https://www.esv.se/statsliggaren/regleringsbrev/?RBID=20264 (Accessed Jan 19, 2022; our translation) We have translated "enhetliga" to "consistent", although "uniform" might be a more literal candidate. We use "statistical assessment support tool" interchangeably with "decision-support tool".

unemployment and the statistically estimated probability of finding a job, produce one of the three outcomes mentioned above (see Fig. 1). For cases that are near a decision threshold, randomization sometimes flips the decision. The purpose of randomization is to enable the effects of interventions to be studied [40].

The statistical model is a neural network trained on historical data consisting of 1.1 million profiles collected over a period of 10 years. Factors considered by the model pertain 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 [40]. The thresholds between different outcomes are subject to political or administrative decisions related to e.g., available resources and volume goals.

Decisions concerning Prepare and Match are formally made by caseworkers, who are instructed in guidelines to primarily adhere to the automated recommendation; overriding a negative recommendation from the system (outcome 1 or 3) requires contacting a special working group within the agency [42]. The central role given to the AI in the decision-making process for Prepare and Match highlights the significant impact of AI for jobseekers. Even if a jobseeker is not offered access to Prepare and Match and may get other support from the agency, the agency itself acknowledges the potential impact of the automated recommendation, since decisions about employment support may affect a jobseeker's possibilities of finding a job [42].

Decisions 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 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 [43], which caseworkers are instructed not to change.[5] Below is an example of a positive decision, i.e. outcome 2 above (our translation):

> By comparing your information with statistics, we have tried to assess how near you are to 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.

Jobseekers can turn down the offer to enroll in Prepare and Match, although this could jeopardize their right to compensation (Swedish Unemployment Insurance Act 1997:238). Furthermore, if jobseekers inform the agency that they will no longer be unemployed within 3 months, the caseworker should not offer the initiative, regardless of what the system recommends [43]. However, as far as we can tell, the jobseeker's own judgment concerning her need for support is not considered by the decision-support system or in the decision-making process at large.

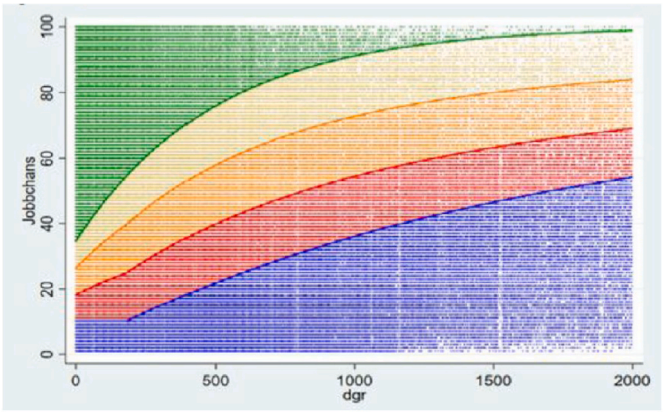

Fig. 1. Relationship between the estimated probability of finding a job ("jobbchans"), number of days of current unemployment ("dgr"), and outcome. The green area (top left) corresponds to "too near the job market" (outcome 1), the blue area (bottom right) to "too far away from the job market" (outcome 3), and the intermediate areas to "suitable for Prepare and Match" (outcome 2 at three different levels, affecting the amount of compensation that providers receive). For example, if the job chance is estimated at 50 % and the unemployment duration is 500 days, the jobseeker is recommended access to Prepare and Match. Note that this illustration is not presented to caseworkers or jobseekers. Also note that thresholds are sometimes adjusted by the agency; the figure conveys the configuration as of December 7, 2021 [40]. Reprinted with permission from the Swedish Public Employment Service [41].

4.2. Performance

Our analysis of trustworthiness begins with performance, the first criterion listed in Table 1. When the Swedish PES measured accuracy as the fraction of historical data points for which a correct decision is recommended, the result was 68 %, which the agency describes as "relatively high" [44]. In this context, a decision is considered correct if the system recommends outcome 1 (too near the job market) and the jobseeker turned out to be employed after 6 months, or if the system recommends outcome 2 (suitable for Prepare and Match) or 3 (too far away from the job market) if the jobseeker turned out to be unemployed after 6 months. This accuracy can be compared with a random baseline of 57 %. It can also be compared with a hypothetical system that deems all jobseekers as needing support (corresponding to outcome 2 or 3); such a system would have an accuracy of 82 %. This can be explained by the fact that most registered jobseekers in fact do not find employment within 6 months, and by the fact that the deployed system is partly governed by thresholds related to volume goals and budget restrictions.

The reported performance differs across sub-populations, as detailed in Table 2. The group "jobseekers with disabilities" has the highest accuracy and true positive rate, but also the highest false positive rate. In contrast, the lowest accuracy and true positive rate is reported for the group "youth", which also has one of the lowest false positive rates. This means that the system is better at correctly classifying disabled jobseekers as needing support than what it is for young jobseekers; however, it also means that it more frequently recommends enrolment in Prepare and Match for disabled jobseekers who do *not* need support than what it does for young jobseekers.

It should be noted that due to the unification of outcomes 2 and 3 in the agency's analysis, the reported performance results only pertain to one aspect of the system's task, namely, to identify whether a jobseeker is too near the job market (does not need support) or is *not* near the job market (needs support). In reality, the system also classifies need for support in two sub-categories: suitable for Prepare and Match (outcome 2) or too far away from the job market (outcome 3). However, as far as we can tell, the agency has not presented any performance results for this aspect of the system's task. Hence, no information is currently available regarding performance for the system as a whole. It should also be emphasized that the reported metrics pertain to a situation where the model is constrained by a particular volume goal; in reality, performance can increase or decrease depending on the agency's adjustment of thresholds.

As for performance comparisons between the system and

[5] Arbetsförmedlingen, Bär för newbies 2.0 (Powerpoint presentation).

Table 2
Performance metrics for a historical population of jobseekers at large as well as for sub-populations [44]. "True positive rate" refers to the fraction of positive cases (jobseekers that turned out to be unemployed after 6 months) that were assessed to need support. "False positive rate" refers to the fraction of negative cases (jobseekers that turned out to be employed after 6 months) that were assessed to need support; we have estimated this metric on the basis of the agency's published metrics.[a]

Population	Accuracy of deployed model	Accuracy for random baseline	True positive rate (TPR)	False positive rate (FPR)
Overall	0.68	0.57		
Men	0.67	0.56	0.67	0.32 ± 0.04
Women	0.68	0.57	0.67	0.26 ± 0.07
Without disabilities	0.64	0.52	0.62	0.28 ± 0.05
With disabilities	0.81	0.80	0.87	0.79 ± 0.03
Born in Sweden	0.61	0.48	0.54	0.20 ± 0.04
Born in another country	0.73	0.67	0.76	0.48 ± 0.05
Not youth	0.70	0.59		
Youth	0.56	0.44	0.47	0.19 ± 0.04
Age 50 years or younger	0.66	0.55		
Older than 50 years	0.71	0.60	0.72	0.32 ± 0.04

[a] The agency's published data does not contain FPR (which is relevant e.g. when assessing fairness). We requested more detailed metrics from the agency but were informed that the original data had not been saved. Therefore, we reconstruct FPR using the published metrics (accuracy, F1 scores and TPR). Since the published metrics are available with only two decimals, the error margin for FPR is large when reconstructing it analytically with interval arithmetic. Instead, we estimate FPR by numerically iterating combinations of confusion matrix values using a coarse-to-fine strategy to identify matrices that reconstruct the published metrics within rounding-error margin (absolute error $\leq$0.005), calculate FPR for each valid matrix and then identify the FPR range.

caseworkers, the available evidence is relatively scarce. No comparison between the deployed model and caseworkers has been published by the agency. However, in one study the agency measured caseworkers' ability to predict how long it will take for a jobseeker to find a job, and compared this with the performance of a simple linear model [45].[6] The comparison showed that for long durations of unemployment, the statistical model had a much higher true positive rate than caseworkers (approx. 60 % compared to 10 %), although also a much higher rate of false positives (approx. 15 % compared to 3 %). No clear differences were obtained for shorter durations of unemployment. The statistical model included in this analysis is reported to perform somewhat worse in terms of overall accuracy (66 %) than the deployed system (68 %), but overall accuracy for caseworkers is not reported.[7] All in all, it is difficult to draw any conclusions about performance differences between the deployed model and caseworkers based on the available evidence.

Furthermore, the reported comparison between model and caseworker concerns the assessment of jobseekers' *need for support* (framed as a prediction of future employment status). Nonetheless, this does not necessarily say much about the *quality of the decisions about employment support* that are made based on such assessments. To study the quality of such decisions, one would also need to consider the *kind of support* that is offered (or not offered). For example, if there are good reasons to believe that enrollment in Prepare and Match would not increase a particular jobseeker's employment chances, but that some other kind of support would, then offering access to Prepare and Match is not a good decision – even if the assessment regarding the jobseeker's overall need of support is accurate.

In line with this, a proper evaluation of the decisions by the system vs caseworker regarding access to employment support would need to consider the *effects* of decisions. While the agency's use of randomization seems to serve the purpose of measuring such effects, no results of these measurements have been presented as far as we can tell. In other words, even if some evidence would show that the deployed system outperforms caseworkers when it comes to assessing the need for support, this would not necessarily mean that the same is true when it comes to the quality of decisions.

One possible way to improve the system' *performance* is by involving jobseekers more actively in the decision-making process. Today, the system's decision does not involve information about needs, wishes or abilities provided from the jobseeker, but only registered data about the individual and her/his area of residence. Comparing this with when the Danish PES trained a statistical model using various factors, including both administrative data and answers to questions asked to the jobseekers, the model learned that the jobseeker's own assessment about their expected duration of unemployment was the most predictive factor [46], indicating that performance can be substantially improved when jobseekers' expertise about their own situation is leveraged.

In this light, it seems reasonable to somehow include the jobseeker's own assessment when making decisions about employment support. This can be done as in Denmark, i.e., by including jobseekers' assessments as one of the factors analyzed by the predictive model. Alternatively, caseworkers can be encouraged to consider the jobseekers' own assessments when making their decisions, alongside the automated recommendation (and potentially also their own judgment). A more salient role of jobseekers in the decision-making process could potentially also enhance the trustworthiness of the agency by making caseworkers morally more invested in each jobseeker's needs and interests.

A possible reason why jobseekers in Sweden are currently not involved in decisions is the potential conflict with consistency and (cost) efficiency. Jobseeker's own assessment can be perceived to make the decision-making process more subjective and open to manipulation [47]. As one manager at PES states: "In order to get as good an assessment as possible, it should be based on actual data that we actually know and not on what we ask the person" (manager, interview). For example, if a certain decision is economically favorable, then a jobseeker could potentially misuse the agency's trust and misrepresent her need for support to get the desired decision. On the other hand, if the agency wants jobseekers to trust it, some amount of reciprocal trust may be needed.

So far, our discussion about performance has pertained to the accuracy of the system as such and when compared to the performance of caseworkers (Q1a). What is perhaps even more important is to assess whether caseworkers' decisions are *improved by assistance from the system* (Q1b), which we call augmented performance. However, as far as we are aware, the only available evaluation of augmented performance is indirect and inconclusive. The agency studied the effect of adding caseworkers' assessments about the jobseeker's levels of motivation, social competence, and work competence to the simpler statistical model described above, along with assessments about potential measures that would increase a jobseeker's employment chances. This kind of augmentation did not improve the performance of the statistical model [45].

A possible interpretation is that additional factors assessed by caseworkers during interaction with the jobseeker are not predictive of future employment chances. Therefore, one may hypothesize that the quality of decisions does not increase when caseworkers overrule the system on the basis of such additional factors. However, to validate this hypothesis one would need to measure the accuracy of caseworkers'

[6] The model used linear regression to estimate duration of unemployment on the basis of 10 factors; this prediction was then binned into the five categories of unemployment duration between which caseworkers were asked to make a choice.

[7] The report contains visualizations of true and false positive rates for five different categories of unemployment duration, but no exact metrics.

decisions with and without decision support [19], which, to our best knowledge, has not been done. In other words, Q1b cannot be answered at this point.

When it comes to information about performance towards stakeholders (Q1c), we observe that neither the overall accuracy of the system nor accuracy for sub-populations and decisions is communicated to caseworkers (e.g., in the case management system) or to jobseekers (e.g., in decision letters), despite the fact that accuracy is far from perfect and varies greatly across different sub-populations. Arguably, this makes it difficult for both caseworkers and jobseekers to assign adequate degrees of reliance on the system.

4.3. Calibration

The system's estimated probabilities of a jobseeker's future employment status are communicated to neither caseworkers nor jobseekers. (In other words, Q2a is answered negatively.) This lack of information makes it difficult for caseworkers and jobseekers to assess whether specific automated assessments can be relied on.

For example, if the model predicts a job chance of 5 % for jobseeker A and 45 % 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. From the perspective of the caseworker, the difference in certainty between cases such as A (high confidence) and B (low confidence) is especially important in situations where the caseworker disagrees with the automated recommendation. If the caseworker disagrees with the automated recommendation and the system is very unconfident, overriding the system may be more warranted than if the system is very confident. Similarly, from the perspective of jobseekers, understanding whether a decision is considered straightforward and univocal, or a borderline case with high uncertainty, can help the jobseeker assess if there is room for negotiation or appeal.

To increase reliability, the statistically estimated job chance could be communicated to stakeholders. However, such a mitigation would raise the question of whether the probability estimates themselves are reliable. As far as we are aware, no evaluation of the studied system's degree of calibration has been published, and without access to the model, it is impossible for researchers to perform such an analysis. (In other words, the answer to Q2b is unknown.) Given the significance of having a well-calibrated system and the importance of stakeholders having access to information about the confidence, this constitutes a major flaw in terms of trustworthy AI.

4.4. Interpretability and explainability

Judging by formal properties of the system, *interpretability* of the AI tool is weak. This is due to the fact that the system uses a neural network which processes information in a non-linear way with large amounts of interactions between variables [23]. Even for AI experts with full access to the model, it is generally difficult to understand how this kind of model reaches its judgments. This means that the inner workings of the system are incomprehensible for stakeholders, i.e., the answer to Q3a is negative.

In order to achieve some degree of *explainability*, the PES uses one of the most popular techniques to approximate the logic of the neural network, called LIME [25]. While LIME and similar methods can give some insight into how an opaque model operates, the methods have been shown to be unstable, reflected in the fact that different explanations can be generated for the same prediction. Furthermore, since methods of this kind are approximate, explanations are not always faithful with respect to the outcomes that they are supposed to explain [26]. For example, if the statistical model predicts that a jobseeker will be *un*employed within 6 months, and the system presents some factors as the most important, then the presented list of factors may sometimes explain why the model predicts that the jobseeker will be *employed* within 6 months (which it did not). For these reasons, Q3b receives a negative answer based on previously documented weaknesses of the explanation method.

In addition to issues regarding inconsistency and unfaithfulness associated with the explanation method as such, an additional concern is raised by the special treatment of one of the factors, namely unemployment duration. 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 coded 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 Fig. 1, the effect of unemployment duration on decisions diminishes as duration increases. For jobseekers that have been unemployed for a long time, the job chance (estimated on the basis of various factors) arguably has more impact on decisions than unemployment duration. Consequently, always mentioning unemployment duration as the most important factor seems potentially inaccurate.

To illustrate the potential benefits of a more interpretable model, we can consider the Danish PES and its use of 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 an 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 categorize the risk of unemployment into three different probabilities [46].

Judging by its formal properties (sparsity and few interactions), the internal logic of the Danish model is more interpretable than the Swedish one. 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 contrastive reasoning, since it is easy to see how an alternative path leads to a different outcome. One could potentially also generate explanations in natural language automatically, as e.g., "Since you are unconfident about finding a job, your statistically estimated job chance is quite low", easily incorporating many of the ingredients found in human explanations (see section 4.5).

The Swedish PES has also experimented with two models that are much simpler and interpretable than the deployed one: a linear regression model, and a combination of a decision tree and 6 linear regressors, both of which predict the expected duration of unemployment (rather than the probability of becoming employed within a certain timeframe). Unlike neural networks, linear regressors treat variables independently and monotonically, making it easier to explain their logic [23]. However, unlike decision trees, their outcomes always depend on all factors, with a less immediate connection between variables and outcomes.

To faithfully explain a prediction made by a linear regressor in natural language, while trying to keep things simple, a decision statement such as the following can be conceived[8]:

> Through statistical profiling, we have tried to assess your need for support. Our assessment is that you will get the best help from a supervisor at one of the providers within the initiative Prepare and Match. On a range from 0 to 10, your need for support is estimated to be **4** (where a higher score indicates a more substantial need for support). Prepare and Match is offered in the range 3–7.
>
> The main circumstances that are deemed to **decrease** your need for support are the beneficial labor-market conditions where you live (−3 points). The main circumstances that are deemed to **increase** your need for support are your relatively long duration and history of

[8] Note that the example is hypothetical and has been purposely designed to demonstrate potential gains in intelligibility. In reality, even simple linear models can be difficult to interpret if the factors correlate with each other.

unemployment (+5 points) and the fact that you seek a part-time occupation (+2 points).

As the example shows, this kind of explanation not only ranks factors by importance but also clearly shows *how* the different factors contribute to a specific assessment.

Importantly, the enhanced interpretability eliminates the need for an approximate explanation method such as LIME, regardless of whether e.g., a decision tree or linear regressor is used. Furthermore, the relative simplicity of the models does not necessarily seem to decrease accuracy. The simplest of the alternative models tested by the Swedish PES has an accuracy of 66 %, compared with 68 % for the deployed model [45]; the slightly more sophisticated one has an accuracy of 74 % [48], i.e. *better* than the deployed model. This suggests that a simpler model can fulfill the stated goals – consistency and accuracy – equally well, or even better, than an opaque model, with substantial gains in intelligibility.

4.5. Availability and intelligibility

When it comes to *availability* of relevant information, PES fairs less than well on this criterion, as revealed by qualitatively comparing rationales for decisions with information about the actual decision-making logic. Outcomes from the system are based on output from a statistical model, the jobseeker's current duration of unemployment, and thresholds that are continuously adjusted by the agency. However, neither the overall nature of this decision-making logic nor how it plays out for specific decisions is communicated to caseworkers or jobseekers. Specifically, stakeholders are not informed about the estimated job chance and how it affects a decision. Furthermore, the existence of decision thresholds is mentioned neither in explanations for specific decisions (in statements to jobseekers or in the case management system), in general information to the public on the agency's website, or in any of the caseworker manuals that we have studied.[9] The lack of transparency also concerns the internal logic of the statistical model. Hence, Q4a is answered negatively.

Arguably, concealing some of the factors that underpin decisions impedes caseworkers' and jobseekers' ability to understand the basis for the decisions recommended by the system. The lack of transparency makes it difficult to assign adequate degrees of reliance concerning decisions recommended by the system and to assess if there is room for negotiation or appeal since the distance to the decision threshold is not communicated. Furthermore, the concealed effect of thresholds introduces unpredictability in the decision-making process, especially for jobseekers who are affected by changes in thresholds. For example, if the agency increases the threshold for positive decisions, some jobseekers may obtain a negative decision as a direct consequence of the changed threshold, without receiving any information about the impact of the changed threshold on their decision.

A possible reason for the agency's choice not to disclose information about the estimated job chance is the risk that jobseekers are discouraged when the estimated chances are very low [47]. As a remedy, one may consider using a carefully selected choice of categories (e.g., substantial/moderate/limited need of support) rather than a percentage, or provide details only when requested by the jobseeker.

Another possible reason for not disclosing the estimated job chance is that the notion of "job chance" is difficult to comprehend given the nature of the statistical model. Since the model does not consider current unemployment duration, which is the most predictive factor, "job chance" becomes a somewhat misleading and unintuitive concept, akin to using the term "dementia risk" with respect to a model that predicts risk for developing dementia without considering the patient's age. A more precise term such as "job chance when disregarding unemployment duration" would be less misleading but equally unintuitive. The challenge is not linguistic, but conceptual. Instead of using an inherently enigmatic notion, the problem could be addressed at its root by eliminating the special treatment of current unemployment duration altogether and including the factor in the statistical modeling instead. With such a design, the model estimates job chance, period. This would arguably make it easier for stakeholders to conceptually grasp the function of the statistical model and thereby also the basis for the system's assessments. (This suggestion is in line with Principle 1 in Ref. [23] which states that for a model to be easily understood by humans, it should obey domain-specific constraints.)

A conceptual simplification of the decision-making logic would also make it easier to communicate decision thresholds in a comprehensible manner. If the agency would decide to present the thresholds with the existing decision-making logic, they would probably need to do it graphically, as in Fig. 1, which may be difficult for laypersons to understand. By eliminating the special treatment of one of the factors, "job chance" can instead be presented along a single probability scale, allowing the thresholds to be marked along the same scale. For example, an explanation of an automated recommendation could state that since the jobseeker's estimated job chance is 60 %, and Prepare and Match is currently offered in the range between 15 % and 80 %, the jobseeker is offered Prepare and Match. By clearly indicating how much the estimated job chance would need to increase or decrease to yield another outcome, the explanation is also contrastive. To further enhance transparency, historical threshold adjustments could also be communicated, at least when they affect a decision.

The ability to conceive the distance to the decision thresholds makes it possible for jobseekers to assess if there is room for negotiation and chance for appeal. The same can be said for caseworkers in relation to the special working group who can permit exceptions from the principle to generally follow the system's recommendation.

Since stakeholders have better reasons to rely on assessments in cases where accuracy is high, and since the system's accuracy varies substantially between groups, it would also be relevant for stakeholders to be informed about the system's accuracy. For example, due to the low accuracy for young jobseekers (see section 4.2), the case management system could generally supply recommendations concerning young jobseekers with a warning concerning the system's low accuracy for this group.

Regarding *intelligibility*, jobseekers' experience as subjects of AI-assisted decisions has not been studied as far as we know. Therefore, we cannot directly assess how they perceive the intelligibility of the explanations for decisions about employment support. Nevertheless, such data exists when it comes to caseworkers and agency officials. In an external study, difficulties in understanding the basis for the system's judgments were voiced, hence Q4b receives a negative answer. The difficulties in understanding are here exemplified by the following response from a caseworker ([40], our translation):

> we discuss a lot because it may happen that I get no on a client, or vice versa, and one is like how does the system work? One cannot really figure out why one got that decision. […] sometimes one cannot really understand what tilts the scale if you get a no for example […]

In our own interviews, some interviewees describe the tool as a "black box", as exemplified by the following excerpts:

> This black box … you do not really know why it has come up with a certain decision. […] I think instructions to officials say that they can attempt to explain what parameters are entered into this machine. However, they still cannot explain exactly why this decision was

[9] This does not rule out the possibility that caseworkers are informed about the thresholds via some other information channel. The only document that describes the actual decision-making logic, as far as we are aware, is a report that contains a figure illustrating the relationship between estimated job chance, unemployment duration and outcome [41]. Even in this case, however, no formalization of the logic is presented.

made in [a specific] case, because the decision is based on other job seekers. (Qualified strategist)

It feels like there is a lot going on behind the scenes and not everyone has access to this information. I absolutely think it is a black box. (Caseworker)

A perceived lack of unintelligibility is also indicated by the fact that the agency provides caseworkers with additional information about how to understand the workings of the statistical model. In an internal document, various examples of "how the model works" are provided (our translation):

If a jobseeker has a lower level of education, the statistical model will take this into account and assess that he/she is somewhat further away from the job market.
[…]
if you live in a commune with high levels of unemployment, you will be assessed to be further away from the job market than if you live in a commune with lower levels of unemployment.

The very existence of this document suggests that caseworkers do not find the list of most important factors to constitute a satisfactory explanation for an automated recommendation. The content of the document also says something about what it is that caseworkers need in terms of explanations, but that is currently lacking in the system. One such ingredient is a *clear connection between factors and outcomes*. In the examples provided in the internal document, connections between factor and outcome are monotonic: when the value of a factor increases, the distance to the job market either increases or decreases, depending on the factor. Furthermore, the provided examples treat factors as being *independently* related to the outcome: a higher level of education decreases the distance to the job market, *regardless of other factors*. Presumably, the monotonicity of the general connections and the independent influence of factors on outcome are two of the ingredients that make this kind of explanation satisfactorily *simple* [27].

Another way to assess the intelligibility of the system's explanations is to compare them with human explanations not involving AI. In caseworker guidelines, the Swedish Social Insurance Agency provides the following example of a decision statement:

Your sickness benefit is decreased from one half to one fourth's compensation. The reason is that you have worked 24 hours per week during […] Considering that you have been able to work more than half time, the Social Insurance Agency deems that your work capability is significantly improved. Therefore, you cannot get more than one fourth's compensation. [49], our translation]

This example highlights two other ingredients that the ranking of factors lack, namely the property of being *local* and providing *contrast*. Specifically, the explanation pinpoints the amount of time that the subject has worked (24 h per week) and the threshold relevant for the decision at hand (half time, i.e., 20 h per week). From this statement, the subject can infer that if she/he had worked 20 h per week or less, the sickness benefit would not be decreased.

One can also observe that the explanation is expressed in natural language (in this case as a decision followed by a line of reasoning). In contrast, the automated explanations used by the Swedish PES merely present factors as a numbered list. All things considered, these comparisons indicate that the explanations currently provided for the automated recommendations lack many of the ingredients – simple logical structure, locality, contrast, and natural language – found in analogous human explanations, which could potentially illuminate why the automated explanations have been found to be unintelligible.

To make the model more intelligible for jobseekers and caseworkers, they can be given the ability to ask "what-if" questions to the system. This way they could explore how potential changes in circumstances affect a decision. For example, if a jobseeker is deemed to have a low chance of finding a job, getting answers to questions such as "What if I move to Stockholm?" or "What if I get a university degree?" could potentially enable jobseekers not only to get advice on how to increase their job-finding prospects, but also support conditions for appealing unfavorable decisions. Supporting these kinds of hypothetical questions is technically trivial and does not require the statistical model to be replaced; users only need to be equipped with a graphical interface that 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?", where the feasibility of changes in circumstances can be addressed in a dialogue between the system and jobseeker [50]. A similar form of interaction could be provided to agency officials, potentially enabling them to better understand how the system makes its assessments.

4.6. Equal and fair treatment

The statistical model that estimates the probability of being employed 6 months into the future is consistent in the sense that the neural network always produces the same output for a given input. However, as for the decision-making system as a whole (within which the statistical model is only one component), two sources of inconsistency can be observed. First, cases near a decision threshold sometimes receive an inverted decision due to the use of randomization. Second, since the agency sometimes adjusts threshold levels, decisions are made in a way that is somewhat inconsistent over time. Hence, Q5a is answered negatively.

It may be argued that to the extent that inconsistencies in decision-making serve a purpose, they can potentially be justified. For example, randomization enables the effects of decisions to be studied systematically, which is clearly important from the perspective of the agency's overall ability to help jobseekers. Furthermore, the adjustment of thresholds can be motivated by an ambition to use the agency's resources optimally. Nevertheless, both types of inconsistencies are a problem from the perspective of equal treatment.

In a comparison with human decision-making, the aspect of the decision-making process that seems relevant to compare is the prediction of future employment status, since the other aspects (involving adherence to administratively controlled thresholds) are mechanical and do not involve judgements. In light of previous research demonstrating that human judgements and decisions tend to be noisy (see e.g. Ref. [29]), one could argue that statistical models in general are more consistent than humans. However, it should be stressed that more complex models such as deep neural networks are known to be non-robust [51]. In other words, a small change in input can cause a large change in output – a property sometimes referred to as adversarial vulnerability. In principle, a lack of robustness could yield situations where *identical* cases yield *identical* judgements, but *similar* cases yield *dissimilar* judgements. For example, adding a humanly imperceptible layer of noise onto an image of a stocking can cause a generally well-performing model to wrongly classify an image of a stocking as an elephant [52]. Similar kinds of vulnerabilities have been demonstrated for non-visual tasks that are more analogous to employment status prediction, such as credit scoring [53].

While there is no available data regarding the robustness of the studied model, the agency mentions intentional manipulation of input data and protection against adversarial attacks as reasons for not disclosing the model openly, suggesting that the model might indeed be non-robust. To this end, it does not seem evident that the statistical model makes judgements more consistently than caseworkers.

As for fairness, among the definitions listed by Ref. [30], the most relevant for the case at hand is *equalized odds* [54], which in this case means that jobseekers with actual need for support should have a similar

classification, regardless of whether they belong to a protected group; analogously, jobseekers *without* need for support should also be classified similarly [55].[10] For example, jobseekers with similar need for support should have a similar probability of being offered enrolment in Prepare and Match, regardless of whether they have a disability. In terms of performance metrics, equalized odds implies that compared groups should have equal true positive rates as well as equal false positive rates [54]. When applying this definition to the performance metrics in Table 2, we see that the model is fair with respect to sex (similar true and false positive rates for women and men). However, when it comes to age, true positive rate differs substantially across categories, ranging from 0.47 for youth to 0.72 for jobseekers older than 50 years. Unfair treatment is observed also for false positive rates for the respective categories (approx. 0.19 for youth and 0.32 for age>50). In other words, jobseekers with similar need for support are classified differently depending on their age. Specifically, older jobseekers are much more likely to be offered support when they need it, but also when they do *not* need it. As for disability, true positive rate is higher for disabled (0.87) than for non-disabled jobseekers (0.62), meaning that among individuals with an actual need for support, disabled jobseekers are more likely to be offered enrolment. On the other hand, false positive rate is *also* higher for disabled (approx. 0.79) than for non-disabled jobseekers (approx. 0.28), meaning that disabled jobseekers are more likely to be offered support also when they do *not* need it. Still, as mentioned in section 4.2, since the available performance results only pertain to one aspect of the system's task (distinguishing whether a jobseeker needs support), it is currently not possible to assess fairness when the other aspect (distinguishing whether the need for support is "too large") is also taken into account.

4.7. Legality, negotiation and appeal

The imprecision of the system's explanations and their documented unintelligibility raise doubts as to whether the Swedish legal requirement to provide a clarifying statement of reasons is fulfilled (Swedish Administrative Procedure Act, 2017:900, section 32). Furthermore, the fact that the system classifies jobseekers with similar need for support differently depending on whether they have a disability or not is possibly incompatible with the Swedish Discrimination Act (2008:567). Hence, the answer to Q6a concerning legal compliance is plausibly negative.

In the context of public decision-making, it is often considered central that subjects can appeal unfavorable decisions. Several obstacles in this regard have been touched on in previous sections. First, information about the system's accuracy and confidence is not disclosed to jobseekers. Having access to such information would help jobseekers assess if there is room for negotiation and hope for a successful appeal. Second, the overall lack of transparency in the logic that governs decisions recommended by the system and the unintelligibility of explanations for decisions makes it difficult for jobseekers to understand the basis for decisions and therefore also to challenge aspects of the decisions that may be questionable. In other words, the extent to which the studied system enables affected subjects to negotiate or appeal unfavorable decisions (Q6b) is very limited.

4.8. Accountability and human oversight

Several circumstances put caseworkers' degree of accountability into doubt. Caseworkers are instructed to primarily adhere to the system's recommended decision, and overruling a negative recommendation is difficult. Furthermore, rationales for decisions, as stated in formal decision letters, are formulated automatically by the system. Indeed, in interviews conducted by the agency, some caseworkers express frustration over not being in control over cases [56]. In other words, Q7 cannot be answered positively.

This situation reflects the first flaw of human oversight policies identified by Green [57]: human oversight policies are not supported by empirical evidence. Green [57] argues that people are unable to provide reliable oversight of algorithms, as they tend to defer to automated systems, reduce their independent scrutiny, and make erroneous judgments about algorithmic outputs. Green [57] also suggests that human oversight policies create a false sense of security in adopting algorithms and enable vendors and agencies to shirk accountability for algorithmic harms. These arguments are relevant to the case of the Swedish Public Employment Service, as they indicate that human oversight may not be sufficient to ensure the quality and fairness of algorithmic decisions affecting jobseekers ([57], pp. 2–3, 8–9).

4.9. Summarizing and analyzing the results

Upon evaluating the application of trustworthy AI criteria in the Swedish PES's Prepare and Match system, several key insights emerge from our analysis. The results, as presented in Table 3, highlight areas where the system currently falls short of aligning with the principles of trustworthy AI.

Firstly, the system's performance, particularly in terms of accuracy and the effectiveness of human-AI collaboration, remains unclear. This uncertainty raises concerns about the reliability of the AI's judgments and decisions across various levels. Furthermore, the lack of clear communication regarding the system's performance to stakeholders

Table 3
Evaluation of trustworthy AI in the Swedish Public Employment Service.

No.	Criteria	Evaluation Question	Answer
1	**Performance**	a. How accurately does the AI make judgments or decisions on all levels?	Unclear
		b. Do human decision-makers make more accurate decisions with the help of the AI system?	Unclear
		c. Is the system's performance communicated to stakeholders?	No
2	**Calibration**	a. Are confidence estimates communicated to stakeholders?	No
		b. If so, are the confidence estimates well-calibrated?	Unclear
3	**Interpretability and Explainability**	a. Can the decision-making logic in principle be understood by stakeholders?	No
		b. Are explanations faithful with respect to the actual decision-making logic?	No
4	**Intelligibility and Availability**	a. Is the decision-making logic made available to the various stakeholders?	No
		b. Are explanations comprehensible for stakeholders in practice?	No
5	**Equal and Fair Treatment**	a. Does the AI make decisions consistently?	No
		b. Are relevant aspects of fair treatment satisfied?	To some extent
6	**Legality, Negotiation, and Appeal**	a. Does the use and functionality of the AI system comply with the law?	Probably no
		b. To what extent does the AI system enable affected individuals to negotiate or appeal unfavorable decisions?	To a very limited extent
7	**Accountability and Human Oversight**	Are human decision-makers able to oversee the operation of the AI and make independent decisions on the basis of the system's output?	No

[10] We disregard outcome-based notions of fairness since jobseekers' need for support likely differs depending on their group.

poses a significant transparency issue. In terms of calibration, the absence of communicated confidence estimates to stakeholders, and the lack of information regarding the calibration of these estimates are problematic. The system reportedly does not ensure consistent decision-making and fails to satisfy critical aspects of fair treatment. This shortfall could lead to biases and unfair outcomes, which are detrimental in a public service context.

Interpretability, explainability, intelligibility and availability are also areas where the system does not meet the necessary standards. The decision-making logic is not adequately communicated to stakeholders, the fidelity of explanations for the actual decision-making logic is unclear, and the comprehensibility of these explanations is not ensured. This lack of clarity can hinder stakeholders' understanding of the system. Regarding legality, negotiation, and appeal, there is uncertainty about the system's compliance with the law and its capacity to allow affected individuals to negotiate or appeal decisions. Lastly, the absence of effective accountability and human oversight mechanisms indicates a critical gap in the system. The inability of human decision-makers to oversee the AI's operation and independently make decisions based on its outputs undermines the system's credibility and safety.

5. Analyzing the results utilizing the theoretical framework

The findings from our evaluation of the Swedish PES are highly interesting, primarily due to their predominantly negative nature and in the light of Sweden's reputation for upholding one of the highest standards of government quality globally [37]. Such results might suggest that the development of trustworthy AI in the public sector is more challenging than initially expected. Utilizing our theoretical framework spelled out in Section 2.1 we can to some extent understand the causes for the negative result, what can be done to avoid them in the future, but perhaps also how the theories may need to be developed further.

First, according to institutional theory, societal norms, rules, and expectations influence organizational behavior and decision-making processes in the public sector. This perspective is particularly relevant for understanding the Swedish PES adoption and integration of new technologies like AI and BD, influenced by both external pressures and internal dynamics. Since Sweden generally has laws and regulations that enforce transparency, and we saw that some laws might even be flaunted in this case, and there are strong norms in the Swedish society in favor of upholding these laws, it might be strange at the offset of why we get this non-transparent result. There are also strong norms in favor of involving citizens in decision-making processes so the non-involvement might also seem unexpected given what institutional theory should predict.

However, there are additional factors to consider in our case. The adoption of neural networks, which are notably opaque, is relatively new in Swedish bureaucracy. Therefore, even if there are strong general norms favoring transparency and openness, this does not necessarily lead to a norm advocating for the use of interpretable AI. Previously, traditional rule-based AI systems were interpretable, so this norm was not required. It also appears that the management at the Swedish PES may not fully understand the type of algorithm they are dealing with. As a result, the norms around transparency have no impact in this context. Also, there is a tendency in the Swedish AI and digitalization policy to approach AI technology as a neutral tool to use in order to gain efficiency.

Furthermore, economic considerations likely play a significant role here, which is a critical aspect to bear in mind when implementing AI systems in the public sector. The Swedish government exerts considerable pressure on its agencies to enhance economic efficiency, a goal that is often easier to measure and achieve than ensuring trustworthy AI. Specifically, improved cost efficiency was one of the initial arguments presented by the government when directing the agency to develop a statistical assessment support tool. Furthermore, in response to this directive, the agency itself mentions potential cost reductions caused by facilitating, speeding up or fully automating labor-market related assessments [58]. In other words, these pressures may have influenced the agency to reduce caseworkers' professional discretion in favor of a more streamlined mode of operation. Overall, deploying AI instead of human labor can potentially be more cost-effective, making it tempting, given external economic pressures, to implement AI systems before they are fully tested and established as trustworthy. Cost efficiency could potentially also influence choice of statistical model. As pointed out by Rudin [22], interpretable models can require more effort to develop in terms of both computation and human expertise, compared to more opaque models such as neural networks.

The Resource-Based View (RBV) emphasizes the importance of leveraging internal resources, including technological infrastructure, skilled personnel, and organizational knowledge. This approach is vital in understanding how these internal assets are maximized to enhance the capabilities of AI, BD, HAI, and DI&A, thereby enriching public sector decision-making processes. A notable aspect of the PES is its workforce, which is highly skilled and possesses extensive knowledge and experience in assisting individuals with job placements and employment support. These skills are rare and hard to find in society at large, even though numerous services exist for matching people with jobs and educational opportunities. By applying the VRIN criteria — value, rarity, difficulty of imitation, and non-substitutability — we can observe how PES can gain competitive advantages by effectively utilizing organization-specific resources and tools related to BD and AI (cf. [4]).

Consequently, rather than instructing caseworkers to predominantly adhere to automated recommendations and making it challenging to override negative decisions, it might be more effective to encourage them to actively employ their own judgment when evaluating these automated suggestions. This approach could ensure better use of the PES' resources. However, given the potential trade-off with consistency and accuracy, caseworkers can be required to internally motivate decisions in cases where they override the decision-support system even if it is very confident about its recommendation. This to strike a balance between utilizing the resources of the caseworkers and the AI system in producing trustworthy AI. Additionally, there is a need to enhance AI literacy skills at both the management and personnel levels, thereby better equipping them to handle the implementation and use of AI systems more effectively.

In addition, it should be noted that developing interpretable models requires specialized expertise. Often, this expertise is lacking in many organizations, as highlighted by Rudin [22] who states that "many organizations do not have analysts who have the training or expertise to construct interpretable models at all." This underscores the need for the PES to not only rely on their existing workforce's skills but also to invest in developing or acquiring the specific competencies needed for creating interpretable AI models, which are essential for transparent and ethical AI applications.

Lastly, Ambidexterity Theory illuminates the balance between exploiting existing resources and exploring new technological opportunities. Our study suggests that the PES plausibly has put too much weight on the explorative side of this scale, possibly at the expense of adequately exploiting existing resources. This potential imbalance might be partly addressed through the implementation of suggestions made earlier, such as increasing AI literacy and clarifying the guidelines for AI system implementation. By integrating insights from Institutional Theory and the RBV, the PES can be guided towards greater ambidexterity.

Additionally, new organizational strategies might be necessary, including the formation of analytical teams, the appointment of analytical integrators, ensuring data quality, and standardizing access to data sources. These strategies are aimed at fostering a supportive environment for DI&A. Research has emphasized the impact of various structural factors, like organizational capital and creativity, on decision-making processes. This encompasses managing external pressures and cultural influences to develop and adapt internal resources, an area that the PES and other public institutions should potentially focus on more in

the future. In order to be creative you need to encourage teams to experiment and take risks. However, this needs to be within controlled parameters, where conditions for trustworthy AI are clearly spelled out and implemented.

By applying ambidexterity, one can perhaps find a more responsible way forward in the implementation of new technologies, which can make their AI more trustworthy. As it is now, the AI system that was rolled out had severe deficiencies. This might be because of too much focus on the future and too little on the situation here and now. Utilizing ambidexterity theory, it becomes apparent that one should opt for a gradual technology upgrade that directly improves citizen services while building a robust AI system infrastructure for the future. This way it both exploits the current resources while exploring new opportunities. To do this effectively one needs to have regular and effective feedback loops, both internal and external, to ensure that organizations can quickly collect and respond to new information. This includes stakeholder feedback, relevant data, and internal performance monitoring. Of course, this includes having mechanisms for continuous learning and adaptation overall, where insights from both successes and failures are used to improve processes and strategies. Additionally, by continuously monitoring the external environment, organizations can identify and evaluate potential threats and opportunities early on, giving them the ability to quickly adapt to changing conditions.

Even if the PES deploy some of these processes already today, this system has clearly failed as they have rolled out untrustworthy AI in one large high-stakes program, which implies that processes need to be greatly improved.

6. Discussion of results

In this section, we will begin by discussing the theoretical implications of AI integration within the Swedish PES, highlighting the disparity between theoretical ambitions and practical realities. Then, we will present our policy and managerial recommendations, aimed at enhancing AI trustworthiness and effectiveness in public decision-making. Following that, we will compare our findings with earlier studies, emphasizing the unique aspects of our research in a Nordic context and its contribution to bridging theoretical and technical discussions in AI. Finally, we will address the limitations of our study, particularly its focus on the Swedish PES and the evolving nature of AI applications in public sector decision-making.

6.1. Theoretical implications

The case of the Swedish Public Employment Service (PES) illustrates a significant gap between theoretical ambitions of AI integration in public services and practical outcomes. Despite Sweden's push towards digitalization and the PES' advanced application of AI, the findings indicate discrepancies in the effectiveness and trustworthiness of these systems. This suggests a need for technically more grounded theoretical frameworks on trustworthiness that account for the complexities and challenges in implementing AI technologies in public sector contexts. The difficulties encountered by the PES in meeting criteria for trustworthy AI highlight the theoretical complexity of creating AI systems that are not only technically proficient but also ethically sound, transparent, and equitable. This underscores the theoretical notion that trustworthiness in AI involves multifaceted considerations extending beyond mere technical capability.

The study brings to light the importance of stakeholder engagement in the AI implementation process. The lack of involvement of jobseekers in decision-making processes, as well as the challenges in intelligibility and transparency faced by both jobseekers and caseworkers, underscore the need for theories that emphasize user-centric design and stakeholder involvement in AI systems. This need can be fulfilled by developing the theories discussed in this paper or applying other theories such as design theory. The case study also sheds light on the dynamics of human-AI interaction, particularly in decision-making contexts. The findings that caseworkers are encouraged to adhere to AI recommendations, and the difficulties in overriding AI decisions, provide practical insights into how AI systems can influence human decision-making. This has implications for theoretical models of human-AI collaboration and accountability.

The challenges faced in developing the AI system to deliver consistent and fair judgments reveal a theoretical gap in understanding how AI systems can be implemented in complex, real-world settings. This suggests a need for more nuanced theories on AI deployment, especially in public service contexts where fairness and equality are paramount. The difficulties in ensuring transparency and explainability of AI decisions in the PES context emphasize the theoretical challenge of balancing strives towards increased accuracy with the need for understandable and interpretable outcomes. In the studied case, no apparent trade-off between these desiderata have been identified (in line with previous work by Rudin [22]), but the potential existence of tensions remains an open question [59], as well as how to deal with potential tensions of this kind when formulating best practices for trustworthy AI. This indicates a need for further theoretical development in this regard, especially in public sector applications where decisions have significant social impacts.

6.2. Policy and managerial recommendations

Our analysis has led to a set of refined recommendations for enhancing the trustworthiness and effectiveness of AI systems in public decision-making. These recommendations are focused on improving stakeholder engagement, decision-making accuracy, and the overall transparency and interactivity of AI systems.

First, a key strategy to improve the performance of AI systems in public decision-making is by more actively involving the end-users, such as jobseekers, in the process. This approach is grounded in the recognition that jobseekers possess unique insights about their own situations, needs, and abilities, which are often overlooked by AI systems that typically rely solely on registered data. For instance, a Danish statistical tool revealed that jobseekers' own assessments about their expected duration of unemployment were highly predictive, underscoring the value of incorporating their personal insights into the decision-making process.

Therefore, involving jobseekers is not a symbolic gesture; instead, integrating their assessments can be a substantive part of the AI's decision-making algorithm. This can be operationalized by incorporating a mechanism within the AI system where jobseekers can input their own assessments regarding their job market readiness, career aspirations, and other relevant factors. These inputs should then be systematically factored into the AI's decision-making process. Moreover, caseworkers, or similar professionals in other sectors using AI for decision-making, could be encouraged to consider these self-assessments alongside the AI's automated recommendations. This would not only potentially improve the accuracy of decisions but also invest the decision-makers, like caseworkers, more deeply in the unique contexts of each individual they assist. Such an approach could foster a stronger engagement in the jobseekers' needs and interests, thereby enhancing the overall trustworthiness and responsibility for the decisions made by the AI system.

Second, addressing the challenge of integrating AI in decision-making processes involves rethinking the role of professionals like caseworkers, especially in contexts like the Swedish PES. The current trend leans heavily towards decision-making based on objective, data-driven AI models. However, this approach sometimes overlooks the nuanced understanding that professionals bring to table. As highlighted by a manager at PES, decisions based solely on "actual data" ignores the subjective realities of individuals. There is a concern that involving jobseekers' self-assessments could introduce subjectivity and potential manipulation, yet reciprocal trust is crucial for building a relationship

between the agency and jobseekers.

This challenge calls for a balanced approach where caseworkers are empowered to use their professional judgment alongside AI recommendations. Such an approach does not undermine the value of data-driven AI decisions but complements them with professional insight. For instance, caseworkers can be encouraged to incorporate their assessments of a jobseeker's motivation, social competence, and work competence into the decision-making process, adding a human dimension to the AI's analytical capabilities. However, it's important to note that augmenting AI decisions with caseworker input does not automatically guarantee improved decision outcomes. Past studies have shown that such augmentation did not necessarily enhance the performance of AI systems. Therefore, this recommendation entails a cautious and measured integration of professional discretion. Caseworkers should be equipped not just with AI tools, but also with training and guidelines that help them effectively blend their expertise with AI insights. This strategy could lead to what we term "augmented performance" – where the combined strengths of professional judgment and AI analytics are leveraged for more nuanced and effective decision-making.

Third, our findings underscore the importance of transparently communicating the performance and accuracy of AI systems to all stakeholders, including both professionals like caseworkers and end-users such as jobseekers. Currently, there is a notable gap in the dissemination of information about the overall accuracy of these systems, especially when it comes to different sub-populations and specific decision contexts. This lack of transparency hinders stakeholders' ability to adequately gauge the reliability of the AI system and, consequently, affects their trust in it. In the context of the Swedish PES, for example, neither caseworkers nor jobseekers are routinely informed about the AI system's performance metrics. This omission is significant because the accuracy of AI-driven decisions can vary greatly across different groups and scenarios. By not communicating this variability, stakeholders are left without a clear understanding of when and how much to rely on the AI's recommendations.

To address this issue, we propose a systematic approach to sharing detailed performance data with stakeholders. This could involve integrating performance metrics and confidence indicators directly into the tools used by caseworkers, and including simplified, understandable explanations of these metrics in communications with jobseekers. Such an approach would not only improve transparency but also empower stakeholders to make more informed decisions about their reliance on the AI system. Moreover, this recommendation extends beyond merely sharing data; it involves educating stakeholders about how to interpret and use this information. Training sessions, workshops, or explanatory materials could be developed to help caseworkers and jobseekers understand the significance of different performance metrics and how they might impact their decisions or expectations.

Fourth, an integral aspect of fostering trust in AI systems is enhancing their understandability and transparency, particularly concerning the logic behind their decision-making. This is crucial in contexts where AI systems, like those used by the Swedish Public Employment Service, make complex decisions that significantly impact individuals' lives. Currently, the system's estimated probabilities regarding a jobseeker's future employment status are not communicated to caseworkers or jobseekers, leading to a lack of understanding and trust in the AI's assessments.

To deal with this challenge, we propose the development of more interpretable AI models. These models should not only be efficient in their decision-making but also provide comprehensible and reliable explanations for their conclusions. This could involve simplifying the algorithmic structures or using techniques that make the decision-making process more transparent, such as decision trees or rule-based systems that are easier for humans to understand. Moreover, it's essential to make all relevant aspects of the AI's decision-making logic accessible to stakeholders. This includes disclosing factors like the existence and influence of specific thresholds or criteria used in the decision process. Such transparency would allow both caseworkers and jobseekers to better understand the rationale behind AI-driven decisions, contributing to a more trustful and engaging AI-user relationship.

Finally, enhancing interactivity within AI systems represents a significant step towards more user-friendly and trust-inspiring technology. Interactive features, such as the ability to explore "what-if" scenarios, can greatly aid stakeholders in understanding how potential changes in input data or conditions could affect AI outcomes. This form of engagement not only makes the AI system more approachable but also demystifies its operations, providing a tangible way for users to see the impact of their own information on the decision-making process. Such interactivity can be particularly beneficial in public service contexts, where decisions have profound effects on individuals' lives. By allowing stakeholders to interact with the AI system and explore different scenarios, we can foster a deeper understanding and acceptance of AI-driven decisions. This, in turn, could lead to improved decision-making outcomes and a stronger sense of agency among all involved parties.

6.3. Comparison with earlier studies

This paper extends earlier studies in the field by particularly focusing on the trustworthiness of AI in public decision-making. Our study contributes new insights, filling gaps highlighted by recent reviews [4,13]. Unlike many studies predominantly centered on the United States, our analysis is situated in a Nordic context. This distinction is crucial, as Sweden, our case study, has notably stricter transparency laws regarding AI than the United States.

Additionally, the Swedish public sector's relatively greater resources provide a unique perspective, underscoring the relevance of our findings in a different socio-political framework. Previous research on trustworthy AI in public decision-making often splits into two streams: one abstract, discussing theoretical frameworks and governmental strategies for AI, and the other highly technical, focusing on specific aspects such as calibration. Our work bridges these streams. We not only establish a set of criteria for trustworthy AI but also apply them to a real-world case. This dual approach allows for a more nuanced understanding of the practical implications, challenges, and advantages of these criteria in the context of public decision-making.

Lastly, our study also utilizes ambidexterity theory in the context of AI in public decision-making, an approach that has only been undertaken once before, and that time in a literature review and not in a case study, adding a novel dimension to our analysis.

6.4. Limitations of the study

Our study is not without limitations. It focuses on a specific case within the Swedish Public Employment Service (PES), and hence, the findings and recommendations may not be directly applicable to other public sector organizations or contexts. However, considering Sweden's high quality of government and strong norms for transparency and openness, combined with more resources than what governmental agencies in many other countries have, the significant challenges faced by the Swedish PES suggest that agencies in less ideal circumstances might fare even worse. Additionally, the rapidly evolving field of AI and its application in public decision-making necessitates ongoing research and adaptation of our findings. Therefore, while our study provides valuable insights, it should be viewed as a starting point for further exploration and refinement in the field of trustworthy AI in the public sector.

7. Conclusions

In concluding our study on the application of trustworthy AI in the Swedish Public Employment Service (PES), we recognize that

integrating trustworthy AI in public services involves navigating through complex challenges and encountering critical gaps, as evidenced by the experience of the Swedish PES. Our detailed evaluation of the PES' initiative Prepare and Match has unveiled significant shortcomings in adhering to the principles of trustworthy AI, raising serious questions about the system's performance and reliability, and underscoring the need for enhanced transparency, interpretability, and a stakeholder-centric approach in AI deployment.

The system's ambiguous performance, particularly in its accuracy and human-AI collaboration, casts doubt on the reliability of AI-driven decisions. This is exacerbated by the lack of clear communication with stakeholders, a fundamental aspect in building trust and understanding in AI applications. The challenges faced by the system in terms of calibration, consistency, and fair treatment reveal potential biases and unfair outcomes, emphasizing the necessity to embed ethical considerations and fairness in AI design and implementation. Furthermore, the necessity for AI systems to be interpretable, explainable, and intelligible is highlighted by the system's current state. With unclear decision-making logic and inadequately communicated explanations, the system fails to engage stakeholders and build trust. This serves as a crucial reminder that AI systems need to be not only technologically proficient but also accessible and understandable to all users.

The study also illuminates critical concerns regarding the system's adherence to legal standards and its effectiveness in enabling affected individuals to advocate for themselves through negotiation or appeal processes. This aspect is crucial in ensuring that AI systems in public services operate not only with technical efficiency but also within the bounds of legal and ethical frameworks. The current structure of the PES system appears to fall short in providing transparent mechanisms through which individuals can challenge or understand the decisions made about them. This gap in legal compliance and user empowerment is a significant oversight, potentially affecting the fundamental rights of individuals and undermining the ethical foundations of AI application in public services.

Looking forward, it is imperative for public sector organizations like the PES to view these findings as a catalyst for change. They are called upon to develop AI systems that are not only technically proficient but also ethically sound, transparent, and aligned with societal values. This requires a commitment to enhancing AI literacy among management and staff, ensuring the development of AI systems that are interpretable and explainable, and implementing robust accountability and human oversight mechanisms. Furthermore, we encourage other public sector organizations to draw lessons from our study, using our insights and guidelines as a benchmark for their AI integration efforts. By sharing our findings, we aim to foster a collaborative approach in the public sector, ensuring that AI systems across various domains are developed with an eye towards ethical and societal considerations.

Additionally, engaging stakeholders actively in the AI development and evaluation process and continuously monitoring and evaluating AI systems for ethical standards and societal values might be crucial steps in realizing the full potential of trustworthy AI in public services. This ongoing process must include a commitment to research and adaptation, acknowledging that AI technology and societal values are constantly evolving. By establishing a culture of continuous improvement and learning, public sector organizations can stay abreast of emerging challenges and opportunities in AI application, ensuring that their systems remain effective, fair, and aligned with the public interest. In essence, our study not only provides a snapshot of the current state of AI in public services but also serves as a dynamic framework for future development, urging a proactive and responsive approach to AI integration in the public sector.

Funding

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 and Marianne and Marcus Wallenberg foundation grant number 2018. 0116.

CRediT authorship contribution statement

Alexander Berman: Writing – review & editing, Writing – original draft, Visualization, Resources, Project administration, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Karl de Fine Licht:** Writing – review & editing, Visualization, Writing – original draft, Project administration, Methodology, Investigation, Conceptualization. **Vanja Carlsson:** Writing – review & editing, Writing – original draft, Investigation, Data curation, Methodology, Conceptualization, Funding acquisition.

Declarations of competing interest

None.

Data availability

Raw interview data is confidential. Information received from the Swedish Public Employment Service via email can be made available on request.

Acknowledgements

The authors would like to thank the anonymous reviewers and Jean-Philippe Bernardy for helpful and constructive comments and advise.

References

[1] L. Floridi, Establishing the rules for building trustworthy AI, Nat. Mach. Intell. 1 (6) (2019) 261–262.
[2] K. de Fine Licht, J. de Fine Licht, Artificial intelligence, transparency, and public decision-making: why explanations are key when trying to produce perceived legitimacy, AI Soc. 35 (2020) 917–926.
[3] D. Kaur, S. Uslu, K.J. Rittichier, A. Durresi, Trustworthy artificial intelligence: a review, ACM Comput. Surv. 55 (2) (2022) 1–38.
[4] A. Di Vaio, R. Hassan, C. Alavoine, Data intelligence and analytics: a bibliometric analysis of human–Artificial intelligence in public sector decision-making effectiveness, Technol. Forecast. Soc. Change 174 (2022) 121201.
[5] High-Level Expert Group on AI (AI HLEG), Ethics Guidelines for Trustworthy Artificial Intelligence, 2019.
[6] K. Bærøe, A. Miyata-Sturm, E. Henden, How to achieve trustworthy artificial intelligence for health, Bull. World Health Organ. 98 (4) (2020) 257.
[7] S. Schmid, T. Riebe, C. Reuter, Dual-use and trustworthy? A mixed methods analysis of AI diffusion between civilian and defense R&D, Sci. Eng. Ethics 28 (2) (2022) 12.
[8] J. Huang, P. Beling, L. Freeman, Y. Zeng, Trustworthy AI for digital engineering transformation, J. Integrated Des. Process Sci. 25 (1) (2021) 1–7.
[9] A.B. Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. García, S. Gil-López, D. Molina, R. Benjamins, Explainable Artificial Intelligence (XAI): concepts, taxonomies, opportunities and challenges toward responsible AI, Inf. Fusion 58 (2020) 82–115.
[10] R. Procter, P. Tolmie, M. Rouncefield, Holding AI to account: challenges for the delivery of trustworthy AI in healthcare, ACM Trans. Comput. Hum. Interact. 30 (2) (2023) 1–34.
[11] V. Eubanks, Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor, 2018. St. Martin's Press.
[12] R.V. Zicari, J. Amann, F. Bruneault, M. Coffee, B. Düdder, E. Hickman, A. Gallucci, T.K. Gilbert, T. Hagendorff, I. van Halem, How to Assess Trustworthy AI in Practice, 2022 arXiv preprint arXiv:2206.09887.
[13] R. Madan, M. Ashok, AI adoption and diffusion in public administration: a systematic literature review and future research agenda, Govern. Inf. Q. 40 (1) (2023) 101774.
[14] P.J. DiMaggio, W.W. Powell, The iron cage revisited: institutional isomorphism and collective rationality in organizational fields, Am. Socio. Rev. (1983) 147–160.
[15] J. Barney, Firm resources and Sustained competitive advantage, J. Manag. 17 (1) (1991) 99–120.
[16] N. Turner, L. Lee-Kelley, Unpacking the theory on ambidexterity: an illustrative case on the managerial architectures, mechanisms and dynamics, Manag. Learn. 44 (2) (2013) 179–196.
[17] T. Hagendorff, The ethics of AI ethics: an evaluation of guidelines, Minds Mach. 30 (1) (2020) 99–120.
[18] S. Chowdhury, P. Budhwar, P.K. Dey, S. Joel-Edgar, A. Abadie, AI-employee collaboration and business performance: integrating knowledge-based view, socio-

technical systems and organisational socialisation framework, J. Bus. Res. 144 (2022) 31–49.
[19] F. Cabitza, A. Campagner, R. Angius, C. Natali, C. Reverberi, AI Shall have No Dominion: on how to measure technology Dominance in AI-supported human decision-making, in: Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems, 2023, pp. 1–20.
[20] R. Tomsett, A. Preece, D. Braines, F. Cerutti, S. Chakraborty, M. Srivastava, G. Pearson, L. Kaplan, Rapid trust calibration through interpretable and uncertainty-aware AI, Patterns 1 (4) (2020).
[21] Y. Zhang, Q.V. Liao, R.K. Bellamy, Effect of confidence and explanation on accuracy and trust calibration in AI-assisted decision making. Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, 2020, pp. 295–305.
[22] C. Rudin, Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead, Nat. Mach. Intell. 1 (5) (2019) 206–215.
[23] C. Rudin, C. Chen, Z. Chen, H. Huang, L. Semenova, C. Zhong, Interpretable machine learning: fundamental principles and 10 grand challenges, Stat. Surv. 16 (2022) 1–85.
[24] F. Doshi-Velez, B. Kim, Towards a Rigorous Science of Interpretable Machine Learning, 2017 arXiv preprint arXiv:1702.08608.
[25] M.T. Ribeiro, S. Singh, C. Guestrin, "Why should I trust you?" Explaining the predictions of any classifier. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016, pp. 1135–1144.
[26] E. Amparore, A. Perotti, P. Bajardi, To trust or not to trust an explanation: using LEAF to evaluate local linear XAI methods, PeerJ Computer Science 7 (2021) e479.
[27] T. Miller, Explanation in artificial intelligence: insights from the social sciences, Artif. Intell. 267 (2019) 1–38.
[28] C. Dwork, M. Hardt, T. Pitassi, O. Reingold, R. Zemel, Fairness through awareness. Proceedings of the 3rd Innovations in Theoretical Computer Science Conference, Association for Computing Machinery, Cambridge, Massachusetts, 2012, pp. 214–226.
[29] D. Kahneman, O. Sibony, C.R. Sunstein, Noise: a flaw in human judgment, Hachette UK (2021).
[30] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, A. Galstyan, A survey on bias and fairness in machine learning, ACM Comput. Surv. 54 (6) (2021) 1–35.
[31] J. Kleinberg, S. Mullainathan, M. Raghavan, Inherent Trade-Offs in the Fair Determination of Risk Scores, 2016 arXiv preprint arXiv:1609.05807.
[32] S. Bertea, Towards a new paradigm of legal certainty, Legisprudence 2 (1) (2008) 25–45.
[33] S. Wachter, B. Mittelstadt, C. Russell, Counterfactual explanations without opening the black box: automated decisions and the GDPR, Harv, JL & Tech. 31 (2017) 841.
[34] G. Sartor, F. Lagioia, The Impact of the General Data Protection Regulation (GDPR) on Artificial Intelligence, 2020.
[35] R.K. Yin, Case Study Research: Design and Methods, Sage, 2009.
[36] Ministry of Enterprise and Innovation, För ett hållbart digitaliserat Sverige – en digitaliseringsstrategi, 2017.
[37] S. Dahlberg, A. Sundström, S. Holmberg, B. Rothstein, N. Alvarado Pachon, C. M. Dalli, The quality of government basic Dataset, version Jan22, in: University of Gothenburg, The Quality of Government Institute, 2022.
[38] V. Carlsson, Legal certainty in automated decision-making in welfare services, Public Policy and Administration (2023).
[39] A. Berman, Why Does the AI Say That I Am Too Far Away From the Job Market?, in: Proceedings of the Weizenbaum Conference "AI, Big Data Social Media, and People on the Move, Berlin, 2023.
[40] H. Bennmarker, M. Lundin, T. Mörtlund, K. Sibbmark, M. Söderström, J. Vikström, Krom – erfarenheter från en ny matchningstjänst med fristående leverantörer inom arbetsmarknadspolitiken, IFAU (Institute for Evaluation of Labour Market and Education Policy), 2021.
[41] Arbetsförmedlingen, Rusta Och Matcha. Om Tjänstens Utformning, 2020. Af-2019/0038 1388.
[42] Arbetsförmedlingen, Beskrivning av arbetsmarknadspolitisk bedömning med ett statistiskt bedömningsstöd i Kundval Rusta och Matcha 1.0, 2020.
[43] Arbetsförmedlingen, Arbetsförmedlingens handläggarstöd. Dnr Af-2020/0016 7459, 2020.
[44] A. Böhlmark, T. Lundström, P. Ornstein, Träffsäkerhet och likabehandling vid automatiserade anvisningar inom Rusta och matcha. En kvalitetsgranskning, Arbetsförmedlingen analys 9 (2021) 2021.
[45] P. Ornstein, H. Thunström, Träffsäkerhet i bedömningen av arbetssökande. En jämförelse av arbetsförmedlare och ett statistiskt bedömningsverktyg, Arbetsförmedlingen analys 7 (2021) 2021.
[46] Styrelsen for Arbejdsmarked og Rekruttering, Beskrivelse Af Profilafklaringsværktøjet Til Dagpengemodtagere, 2020.
[47] S. Desiere, K. Langenbucher, L. Struyven, Statistical profiling in public employment services: an international comparison, in: Employment and Migration Working Papers, OECD Social, 2019.
[48] P. Helgesson, P. Ornstein, 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 8 (2021) 2021.
[49] Försäkringskassan, Att skriva beslut i Försäkringskassan, Riktlinje 14 (2005) version 5. Dnr 76321-2010, 2005.
[50] A. Berman, E. Breitholtz, C. Howes, J.-P. Bernardy, Explaining Predictions with Enthymematic Counterfactuals, 1st Workshop on Bias, Ethical AI, Explainability and the Role of Logic and Logic Programming, University of Udine, Udine, Italy, 2022. BEWARE-22, co-located with AIxIA 2022.
[51] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, R. Fergus, Intriguing Properties of Neural Networks, International Conference on Learning Representations, 2014.
[52] S.-M. Moosavi-Dezfooli, A. Fawzi, O. Fawzi, P. Frossard, Universal adversarial perturbations. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 1765–1773.
[53] V. Ballet, X. Renard, J. Aigrain, T. Laugel, P. Frossard, M. Detyniecki, Imperceptible Adversarial Attacks on Tabular Data, 2019 arXiv preprint arXiv:1911.03274.
[54] M. Hardt, E. Price, N. Srebro, Equality of opportunity in supervised learning, Adv. Neural Inf. Process. Syst. 29 (2016).
[55] S. Verma, J. Rubin, Fairness definitions explained, in: Proceedings of the International Workshop on Software Fairness, 2018, pp. 1–7.
[56] E. Hansson, G. Luigetti, Minirapport intervjuer med medarbetare. ESF-projekt Kundval rusta och matcha, Arbetsförmedlingen (2022). Dnr Af-2022/0026 7090.
[57] B. Green, The flaws of policies requiring human oversight of government algorithms, Computer Law & Security Review 45 (2022) 105681.
[58] Arbetsförmedlingen, Förbereda för reformeringen av myndigheten: Återrapport regleringsbrev 2020, 2020. Dnr Af-2020/0034 2118.
[59] G.K. Dziugaite, S. Ben-David, D.M. Roy, Enforcing Interpretability and its Statistical Impacts: Trade-Offs between Accuracy and Interpretability, 2020 13764 arXiv preprint arXiv:2010.