INSTITUTE FOR EMPLOYMENT RESEARCH
The Research Institute of the Federal Employment Agency

IAB-DISCUSSION PAPER

Articles on labour market issues

12|2025

Ethical Integration in Public Sector AI.

The Case of Algorithmic Systems in the Public Employment Service in Germany

Bernhard Bauer, Sabrina Mühlbauer, Kerstin Schlögl-Flierl, Enzo Weber, Paula Ziethmann

ISSN 2195-2663

Ethical Integration in Public Sector AI. The Case of Algorithmic Systems in the Public Employment Service in Germany

Bernhard Bauer (Center for Responsible AI Technologies, and University of Augsburg, Germany)
Sabrina Mühlbauer (IAB, Germany)
Enzo Weber (IAB, and University of Regensburg, Germany)
Kerstin Schlögl-Flierl (Center for Responsible AI Technologies, and University of Augsburg, Germany)
Paula Ziethmann (IAB, Germany)

Mit der Reihe „IAB-Discussion Paper“ will das Forschungsinstitut der Bundesagentur für Arbeit den Dialog mit der externen Wissenschaft intensivieren. Durch die rasche Verbreitung von Forschungsergebnissen über das Internet soll noch vor Drucklegung Kritik angeregt und Qualität gesichert werden.

The “IAB Discussion Paper” is published by the research institute of the German Federal Employment Agency in order to intensify the dialogue with the scientific community. The prompt publication of the latest research results via the internet intends to stimulate criticism and to ensure research quality at an early stage before printing.

Contents

Abstract

This article addresses the ethical design of artificial intelligence (AI) in the public sector, with a particular focus on Public Employment Services (PES). While AI is increasingly employed to streamline administrative processes and improve service delivery, its application in employment mediation raises fundamental concerns regarding fairness, accountability, and democratic legitimacy. The EU Artificial Intelligence Act has further underscored the urgency of addressing these challenges by classifying employment-related AI systems as high-risk, thereby mandating robust safeguards to prevent discrimination and ensure transparency. The central aim of this study is to examine how ethical and social considerations can be systematically embedded in the development and implementation of public sector AI. Using the German PES as a case study, we introduce the "Embedded Ethics and Social Sciences" approach (EE), which integrates ethical reflection and practitioner involvement from the outset. Qualitative insights from interviews with caseworkers highlight the socio-technical challenges of implementation, particularly the need to reconcile efficiency with citizen trust. Building on these insights, we propose concrete design elements emerging from the integration of ethical and social considerations into system development. In this context, we discuss issues of data ethics and bias, fairness, and the role of explainable AI (XAI). Our analysis demonstrates that this framework not only supports compliance with new regulatory requirements but also strengthens human oversight and agency, and shared decision-making. More broadly, the findings suggest that ethically grounded design can enhance fairness, transparency, and legitimacy across diverse domains of public administration, thereby contributing to more accountable and citizen-centered governance in the digital era.

Zusammenfassung

Dieser Artikel befasst sich mit der ethischen Gestaltung von Künstlicher Intelligenz (KI) im öffentlichen Sektor, wobei der Fokus auf den öffentlichen Arbeitsverwaltungen liegt. Während KI zunehmend zur effizienteren Gestaltung von Verwaltungsprozessen und zur Verbesserung der Dienstleistungserbringung eingesetzt wird, wirft ihre Anwendung in der Arbeitsvermittlung grundlegende Fragen hinsichtlich Fairness, Rechenschaftspflicht und demokratischer Legitimität auf. Das EU-Gesetz zur Künstlichen Intelligenz (EU AI Act) unterstreicht die Dringlichkeit der Bewältigung dieser Herausforderungen, indem es KI-Systeme, die die Arbeitsvermittlung betreffen, als risikoreich einstuft und damit strenge Schutzmaßnahmen vorschreibt, um Diskriminierung zu verhindern und Transparenz zu gewährleisten. Das zentrale Ziel dieser Studie ist es zu untersuchen, wie ethische und soziale Überlegungen systematisch in die Entwicklung und Umsetzung von KI im öffentlichen Sektor eingebunden werden können. Anhand der deutschen öffentlichen Arbeitsverwaltung als Fallstudie stellen wir den Ansatz „Embedded Ethics and Social Sciences" (EE) vor. Dieser Ansatz integriert ethische Überlegungen und den Bezug zur Praxis bereits in die Entwicklung des Modells. Qualitative Erkenntnisse aus Interviews mit

Vermittlungsfachkräften verdeutlichen die soziotechnischen Herausforderungen der Umsetzung, insbesondere die Notwendigkeit, Effizienz mit dem Vertrauen der Bürger:innen in Einklang zu bringen. Auf der Grundlage dieser Erkenntnisse geben wir Empfehlungen für die Gestaltung von KI-Systemen, welche sich aus der Integration ethischer und sozialer Überlegungen in die Systementwicklung ergeben. In diesem Zusammenhang diskutieren wir Fragen der Datenethik und Bias, der Fairness und der Rolle erklärbarer KI (XAI). Unsere Analyse zeigt, dass der EE-Ansatz nicht nur die Einhaltung neuer regulatorischer Anforderungen unterstützt, sondern auch die menschliche Aufsicht, die Handlungsfähigkeit und gemeinsame Entscheidungsfindung stärken kann. So deuten die Ergebnisse darauf hin, dass ein ethisch fundiertes Design Fairness, Transparenz und Legitimität in verschiedenen Bereichen der öffentlichen Verwaltung erhöhen kann und somit zu einer verantwortungsvolleren und bürgernahen Umsetzung im digitalen Zeitalter beiträgt.

JEL classification

C49, J14, J16, J64, J71

Keywords

AI Ethics, Artificial Intelligence, Bias, Explainability, Fairness Public Employment Service, Public Sector

Acknowledgements

We are very grateful to the research coordination (FOKO) department at the IAB, especially Sandra Biermeier, Anja Mißling-Matthes and Christopher Osiander for valuable suggestions and comments.

1 Introduction

AI-based tools are now being explored across various areas of the public sector to manage administrative processes, improve service delivery, and support decision-making in ways that promise greater efficiency, speed, and precision (Bullock, 2019). Alongside these ambitions, however, new concerns have emerged - particularly regarding the ethical and social consequences of deploying algorithmic systems in public institutions (cf. Deutscher Ethikrat, 2023). Technological developments in the public sector must adhere to democratic principles, legal mandates, and the obligation to act in the public interest. This sets a higher standard for legitimacy: systems must not only function reliably but must also be aligned with values such as fairness, transparency, accountability, and social inclusion. These requirements are especially critical when algorithmic tools are used in areas that directly affect people's lives and rights. These concerns are reflected in the EU Artificial Intelligence Act (2024), the world's first legal framework for AI, which classifies various AI-based tools in different areas of the public sector as high-risk systems.

The public employment sector is one such area. Employment agencies are responsible for job placement, benefits administration, and career counselling. In fulfilling these tasks, they engage with a broad and often vulnerable segment of the population. Introducing AI into this context - specifically for job matching – bears great potential for improving individual outcomes, but also raises pressing ethical questions: How are matching decisions generated and justified? What safeguards exist to prevent discrimination? And how are affected individuals informed or included?

These questions form the background of the present study, which addresses the following central research question:

How can public sector AI projects meaningfully integrate ethical and social considerations during development and implementation?

Our analysis and the answer to our research question are grounded in a concrete case study: the design of a prototype for an AI-based tool to support job placement within the German public employment services (PES). This study draws on both conceptual approaches and preliminary empirical insights from our project to reflect on how ethical standards can be translated into practice from the very beginning of system development.

The structure of this article is as follows: Chapter 2 reviews the current state of research on the use of AI in the public sector, with a particular focus on PES. It outlines key ethical frameworks, discusses recent international experiences with AI-based and profiling systems in employment services, and examines approaches to embedding ethical considerations in the development and deployment of such technologies. Chapter 3 introduces our case study on AI in the German PES, beginning with an analysis of the specific ethical challenges arising in this context. In Chapter 3.1, we outline our strategy for addressing these challenges by introducing the Embedded Ethics and Social Sciences approach (EE). Using our case study as an example, we explore how public sector AI projects can meaningfully integrate ethical and social considerations throughout both the development and implementation phases. Chapter 3.1.1 presents initial insights from a qualitative interview study with caseworkers, conducted as part of this strategy to involve

practitioners and gain a deeper understanding of the sociotechnical context in which the system will operate. In Chapter 4 we delve deeper into the ethical and technical dimensions of our findings. This chapter outlines concrete ethical design elements we intend to implement. These include a focused discussion of algorithmic bias and data ethics (4.1), as well as the role of explainability as a precondition for human oversight and trust (4.2). In section 4.3, we situate our work within the broader discourse on fairness, accountability, and stakeholder involvement in the development of AI for the public sector. Chapter 5 concludes with a summary of our contributions and outlines the next phases of the project.

2 Public Sector AI

AI is increasingly shaping the functioning of the public sector and public administrations across a range of domains. From predictive policing (Minocher & Randall, 2020) and fraud detection (Dasgupta et al., 2025) to health care logistics (Masmoudi et al., 2021) and labor market interventions, AI technologies are becoming integral to how public institutions anticipate needs, allocate resources, and make decisions. Under the umbrella term *Public Sector AI*, one can broadly understand the application of algorithmic and data-driven systems within state-led, democratically accountable contexts to perform or support administrative, regulatory, or service-oriented functions. These systems differ substantially (from Large Language Models to Computer Vision), but also in purpose, design, and scope. Yet, they are united by the requirement to operate within legal constraints, respect fundamental rights, and uphold principles of public accountability. (Selten and Klievink, 2024)

To guide the ethical development and implementation of such technologies, a variety of normative frameworks have emerged over the past decade. Within the European context, the **AI4People framework** (Floridi et al., 2018) is particularly influential, calling for AI systems that promote *human dignity,* support *human flourishing*, and are embedded in democratic values. More recently, the **European Union's Artificial Intelligence Act (EU AI Act)**, adopted in 2024, has provided a binding legal framework. The EU AI Act introduces a **risk-based regulatory approach,** categorizing AI systems based on their potential to harm fundamental rights, safety, or the public interest. Systems used in public services are frequently classified as high-risk and therefore subject to stringent requirements regarding transparency, accountability, and oversight[1].

Ethical considerations in Public Sector AI typically revolve around several recurring themes. These include the potential for **discriminatory outcomes, automation bias**, lack of **explainability**, and **data protection risks.** Alon-Barkat and Busuioc (2023), for instance, have shown how human decision-makers in public sector contexts are prone to **selective adherence** to algorithmic advice, particularly when the cognitive load is high or institutional pressures favor deferring responsibility. Similarly, Bullock (2019) has explored how AI may reshape bureaucratic discretion and undermine the balance between rule-following and case-sensitive judgement. He argues that algorithmic systems, by prioritizing consistency and efficiency, risk narrowing the discretionary space available to frontline workers, potentially limiting their ability to respond flexibly to individual circumstances and contextual nuances.

Research by Gesk and Leyer (2022) further underscores the importance of **citizen acceptance**, highlighting that perceptions of fairness, transparency, and perceived competence significantly shape the public's willingness to trust AI in administrative contexts. Trust in Public Sector AI is not merely a technical concern, but a foundational prerequisite for democratic resilience. Without trust, meaningful civic participation becomes difficult to sustain, and public confidence in the legitimacy of bureaucratic institutions and the broader system of democratic checks and balances is eroded. Gräfe et al. (2024) examine the effects of automation on public employees

[1] The implications of this classification for our specific case - a job-matching AI system in the German public employment sector - will be addressed in Chapter 4.

themselves, raising questions about deskilling, shifting task profiles, and changing notions of responsibility within the public workforce.

However, despite these commonalities in ethical concerns in Public Sector AI, Wenzelburger et al. (2024) stress the importance of **contextual factors**, arguing that the social and institutional embedding of AI systems substantially affects their ethical and practical implications. There is considerable variation across domains regarding how ethical and social concerns manifest and how they should be addressed. AI in predictive policing, for example, raises different normative questions than algorithmic triage in health care or AI-supported decision-making in social welfare. The stakes, the populations affected, and the degrees of discretion granted to algorithms vary significantly.

Given this heterogeneity, the remainder of this chapter will focus on the domain most relevant to our study: **public employment services (PES)**. As a key interface between the welfare state and the labor market, PES play a crucial role in job placement, benefit allocation, and labor market integration. Their activity substantially affects job findings of unemployed (Hartl et al., 2021). The introduction of AI systems in this field - particularly for job matching - touches upon issues of fairness, professional discretion, and institutional legitimacy in especially sensitive ways. The following sections review the literature on algorithmic decision-making in PES, and draw out both conceptual frameworks and empirical findings from the application of AI systems in various countries that inform the discussion of our approach to ethically responsible system development in Germany.

2.1 AI in Public Employment Services

Public employment services are key institutions in national labor market systems. Their primary task is to connect people seeking work with available employment opportunities. Although the institutional arrangements differ across countries, the fundamental purpose remains the same: to facilitate job matching through information, counselling, placement, and active labor market policies. PES support individuals not only by helping them find jobs but also by offering training opportunities and career guidance. At the same time, they assist employers in recruiting suitable candidates and in addressing workforce needs.

What sets PES apart from private employment agencies is their public mandate. They operate within legal frameworks and are subject to democratic oversight and political accountability. Their actions must reflect and support broader societal values, including equality, non-discrimination, and inclusion. At the European level, these responsibilities are further underscored by the European PES Network. This network brings together the employment services of the EU Member States, as well as those of Iceland, Liechtenstein, Norway, and representatives of the European Commission. One of the Network's stated goals is to contribute to the implementation of the European Pillar of Social Rights. This framework outlines 20 key principles, grouped into three chapters: (1) equal opportunities and access to the labor market, (2) fair working conditions, and (3) social protection and inclusion. PES are thus not merely operational agencies; they are expected to actively support the realization of social rights across Europe. (European Commission, o. J.-a); (European Commission, o. J.-b)

This broader mission shapes the ethical expectations placed on PES. Matching jobseekers with jobs is not a neutral or purely technical task. Gottwald and Sowa (2019) or Osiander and Steinke (2011) show the perspective of caseworkers and how the placement depends on their behavior. Körtner and Bonoli (2021) show that, naturally, discrimination and inequality exist in different fields of the PES. Job placement involves decisions that can significantly affect people's lives, especially in situations where individuals are already experiencing uncertainty or disadvantage. Ethical concerns in this context may relate to how people are classified or prioritized, how their skills and preferences are considered, or how transparent and fair the processes appear to them.

The literature reveals numerous challenging aspects across various fields of of the PES. However, the growing use of digital technologies - including AI systems - adds new complexity to these existing concerns. As many PES experiment with AI-based tools to improve matching efficiency, the need to reflect on ethical implications becomes more urgent. (Desiere and Struyven, 2021)

2.2 International Experiences with AI and Profiling Systems in PES

The international landscape of AI-based and technologically supported profiling systems in PES reveals both the potential and the risks of such tools. Cross-national comparisons underscore the urgent need to embed ethical and societal considerations from the earliest stages of system development, especially in light of past failures and controversies. While some countries have managed to deploy such systems effectively, others have faced backlash due to issues of transparency, discrimination, and lack of legitimacy - challenges that are deeply tied to the design, implementation, and socio-institutional embedding of these technologies.

The OECD's comparative report on profiling systems in PES (Desiere et al., 2019) provides a useful taxonomy for understanding the variation in approaches. It distinguishes between rule-based profiling systems (e.g., Norway, Poland), statistical models (e.g., Australia, United States), and caseworker-based assessments (e.g., Estonia, Germany). Rule-based profiling classifies jobseekers based on administrative criteria like age or education level. These criteria often target groups with a higher risk of becoming long-term unemployed. Most employment services combine these rules with caseworkers' discretion and with a lot of other assessment tools. Statistical profiling uses a statistical model to predict labor market disadvantage. While rule-based profiling is very strictly bound to predefined rules, statistical profiling uses characteristics of jobseekers more flexibly to map individual patterns. Finally, caseworker-based profiling relies only on caseworkers' discretion. They are supported by quantitative or qualitative tools to assess jobseekers' skills and needs. Thus, caseworker-based profiling is more expensive than statistical profiling. For instance, in Germany, after an interview of one hour, caseworkers must distinguish if a jobseeker is "easy" or "hard-to-integrate". This diversity reflects both technical choices and institutional legacies. In what follows, we focus on four countries of particular relevance to our case: Austria, Poland, Sweden, and Switzerland.

Austria's Arbeitsmarktchancen-Assistenz-System (engl.: Labor market opportunity-assistance-system; AMAS) has received particular attention. Introduced with the stated goals of improving counselling efficiency, enhancing the effectiveness of labor market measures, and reducing discretionary bias, the system assigns unemployed individuals to one of three categories based on their estimated likelihood of labor market reintegration. Those in the mid-range category—between a 25% and 66% probability of finding employment within six months—receive the

majority of resources, based on the assumption that support is most effective for this group. The algorithm incorporates a range of features that, while statistically predictive, have socially discriminatory effects. For example, being female (-0.14), having care responsibilities (-0.15), or having physical or mental health issues (-0.67), all reduce the calculated employment probability. Similarly, being over 50 years old reduces the probability by 0.7. This design has sparked criticism for statistical discrimination, lack of transparency, and stigmatizing effects - issues that are not unique to Austria. (Kolleck & Orwat, 2020)

A similar model was piloted in Poland but was discontinued in 2019 following criticism from the public and experts. The system was accused of infringing on privacy rights, exacerbating existing inequalities, and providing insufficient transparency about how decisions were made. (ibid.)

As Williams et al. (2018) have pointed out, statistical profiling models by design assign group-level characteristics to individuals, which can reinforce systemic disadvantages. The OECD (2019, p. 24) has described this as a form of statistical discrimination that is especially problematic for vulnerable groups such as migrants, older jobseekers, women, and those with lower levels of education. Because these social groups tend to have lower average employment probabilities, individuals are often misclassified, either being denied resources they would benefit from or being deprioritized in the allocation of services.

When we address discrimination in this text, we refer to discrimination that is historically and structurally embedded - thus adopting the more social-scientific interpretation of the term, rather than its technical definition, in which anyone disadvantaged by a system is considered discriminated against. In our (social-scientific) conceptualization, only groups that have been historically and structurally marginalized can be regarded as experiencing discrimination.

In contrast, Sweden demonstrates how profiling models can be deployed in a way that mitigates discriminatory effects through *institutional* design. There, the AI-supported system identifies individuals at the highest risk of long-term unemployment, not to deprioritize them, but rather to direct *additional* resources and early intervention measures toward them. This inversion of the usual logic highlights that the ethical evaluation of profiling systems cannot be based solely on algorithmic outputs but must consider their *socio-technical integration* and the normative purposes they serve.

Another key lesson emerges from countries such as Switzerland, South Korea, and Finland, where profiling systems failed to gain traction due to low acceptance among staff/caseworkers. In Switzerland, for instance, the Statistical Assistance for Programme Selection (SAPS) was never scaled beyond pilot stages, despite some positive effects on labour market integration, because caseworkers doubted its accuracy and refused to use the tool in practice (Arni and Schiprowski 2015). Similar patterns of resistance have been observed elsewhere, where systems were rolled back after implementation due to a lack of trust and perceived relevance (Caswell et al., 2010). The OECD report (Desiere et al., 2019) concludes that stakeholder engagement is critical to the success of profiling tools. Involving both caseworkers and jobseekers in the design, testing, and evaluation of such systems is essential not only for ensuring transparency and fairness but also for building the trust necessary for long-term institutional acceptance. Without this, even ‘technically sound’ systems seem likely to fail.

These international experiences offer valuable insights for the ethical development of AI in PES, highlighting two central dimensions: (1) the importance of addressing statistical discrimination through system design and social policy integration, and (2) the necessity of involving all relevant stakeholders - including users and jobseekers - in the development and implementation process. Both aspects are central to our approach and will be taken up in detail in Chapter 3, where we present the socio-technical framework for our case study. Before that, however, we turn to existing research on how ethical considerations have thus far been integrated into AI systems in Public Sector AI and PES.

2.3 Ethical Integration in Public Sector AI and PES

As discussed previously, the development and deployment of AI systems in the public sector - particularly in PES - can draw on a range of ethical guidelines and normative frameworks. The most prominent of these in the European context is the EU AI Act. Beyond such policy-level frameworks, scholars have explored how more fundamental ethical theories can enrich the normative foundations of AI in the public sector. For instance, Allhutter et al. (2022) point to the potential of integrating the capabilities approach - developed by Amartya Sen and Martha Nussbaum - as a means to engage with the complex questions of values, social change, and normative legitimacy raised by technological transformation in welfare institutions. These theoretical lenses provide critical guidance for aligning AI systems with democratic values and social justice. However, as Allhutter et al. also note, a central challenge remains: translating such high-level ethical considerations into concrete practices within the development and operationalization of AI systems. In practice, it often remains unclear how abstract normative commitments can be meaningfully embedded in algorithmic design and real-world implementation.

Feminist approaches in particular have already made significant progress in this direction, i.e., in integrating ethical and social aspects into the development and design of technologies and the economy. The third gender equality report of the German Federal Government (Yollu-Tok et al., 2021) deals with the question: "What course needs to be set to shape developments in the digital economy in such a way that women and men have equal opportunities for realization?". Various recommendations for action are given to achieve the goal of equality. The report focuses on how digitalization can be shaped in a gender-equitable way and what effects, opportunities and risks the digital transformation has for women and men. Scupola and Mergel (2022) also study the digital transformation of public administrations in Denmark. They do this by identifying five co-production phases and investigating how Denmark created public value through its digital transformation efforts. The authors point out that co-production was key in the success of digital transformation and that, in the case of Denmark, an important instrument in the implementation of the digital transformation of public administration was co-financing.

Based on these findings, we are developing a systematic investigation strategy for our use case that takes all the aforementioned ethical and social aspects into account. One approach that is particularly promising from our perspective is the **Embedded Ethics and Social Sciences (EE) approach**, developed by McLennan et al. (2022), and applied in Germany in particular by the Center for Responsible AI Technologies (Jörg et al., 2023; Schlögl-Flierl & Ziethmann, 2025; Ziethmann et al., 2025). This approach advocates for the **continuous integration of ethical and**

social analysis into the entire research and development process. Rather than treating ethical evaluation as an external or retrospective activity, EE seeks to embed it directly within interdisciplinary collaboration. Drawing on traditions such as **value-sensitive design (Friedman et al., 2002)**, and **parallel ethics research** (Van Der Burg, 2009), EE aims to establish lasting, dialogical relationships between social sciences, ethics, engineering, and system users. It is thereby situated firmly within the tradition of sociotechnical integration research (Fisher et al., 2015; Fisher & Schuurbiers, 2013), which regards ethical and technical development as mutually constitutive and interdependent processes.

The EE framework will serve as the overarching methodological orientation for our project. It provides a structured yet flexible approach to ensure that ethical considerations are not merely stated but actively influence design decisions, deployment strategies, and evaluative procedures throughout the system lifecycle. Crucially, an EE-informed approach also underscores the importance of actively involving both **caseworkers and jobseekers** as key stakeholders - not only through empirical studies but also by integrating insights from existing research on their expectations, reservations, and operational logics.

For instance, **Broecke (2023)** emphasizes that while machine learning tools offer significant potential to improve labour market matching, their success hinges on user readiness and the resolution of persistent issues related to transparency, robustness, and fairness. For this reason, this chapter also turns its attention to existing conceptual and empirical studies regarding the aforementioned target groups.

As the related literature shows, especially in Germany, where a caseworker-based profiling strategy is used in the PES, the caseworkers' perspective plays a key role. Freier and Senghaas (2022) point out that caseworkers act as street-level bureaucrats and investigate how they use enhanced discretion in an innovation project. The authors analyze qualitative interviews and group discussions with placement staff. The responding caseworkers identify with the idea of tailoring their services to meet their clients' needs and reducing pressure. There are several factors that influence their use of discretion, like targeting mechanisms within the organization. The authors point out that discretion in itself is neither good nor bad, but the interplay between discretion, organizational norms and the labor market situation can promote individualized services and lasting employment. Dolata et al. (2020) investigate coping strategies of employment consultants and describe them as active and creative supporters of clients in a difficult situation. The study identifies opportunities for design research that support the consultants' mediator role, rather than their individual activities.

Osiander and Steinke (2011) analyze a reform process within the German Federal Employment Agency (BA) by evaluating the reorganization of the business system and the employment agencies' self-perception. The study is based on the premise that the success of reforms hinges on how street-level bureaucrats at the lower hierarchical levels of the labor administration perceive the fundamental content-related aspects. Caseworkers will only approve reforms if they appear comprehensible and sensible from their point of view. This can also be applied to our use case. Acceptance by caseworkers is a fundamental building block in the design of a recommender system for the PES.

Dietz et al. (forthcoming a) present findings from a survey experiment with German PES caseworkers, exploring how factors such as discretionary authority, professional identity, and

individual attitudes towards AI shape system acceptance. They find that perceived preservation of discretionary power is critical, while framing professional identity in terms of client interaction does not significantly influence willingness to delegate decisions to AI. Acceptance, in turn, correlates strongly with general attitudes toward AI. Another study by Dietz et al. (forthcoming b) describes a conjoint experiment. Caseworkers were presented with hypothetical decision-support systems for selecting further training measures. Systems were varied according to attributes such as decision transparency, binding nature of recommendations, source of the software, and time-efficiency. The results suggest a strong preference for internal (non-commercial) systems, non-binding recommendations, comprehensible outputs, and clear time-saving benefits. These findings provide concrete design criteria for aligning AI systems with caseworker expectations and institutional values. Van den Berg et al. (2024) analyze how a ML algorithm, the caseworkers and the jobseekers themselves assess their probability for getting re-employed. They find that the algorithm performs better in terms of accuracy than predictors based on the assessment of caseworkers or jobseekers alone. On average, jobseekers are more optimistic than the algorithm, while caseworkers are more pessimistic. Finally, combining caseworker assessments with ML algorithms could improve the performance.

In sum, these conceptual and empirical contributions offer valuable insights into the conditions under which ethical principles can be practically integrated into the development and deployment of AI systems in the public sector, as well as and the scope pf these principles. They reveal that ethics must not be external to system design, but co-produced through inclusive, interdisciplinary, and iterative processes. In the following chapter, we elaborate this further by closely examining our own case study of AI development in the German PES context, applying the EE approach to explore how ethical design can be achieved in practice.

3 Our Case Study: AI in Public Employment Services in Germany

As outlined in Chapter 2, PESs differ due to historical, political and socio-economic reasons. In Germany, the Federal Employment Agency (BA) is the central operator. The BA is responsible for job placement, job advice, and job promotion, as well as administering German unemployment insurance and regulating unemployment benefits. There, large administrative data on the whole work force in Germany is saved. This extensive data infrastructure enables the development of algorithmic systems capable of supporting job placement decisions at scale. However, it also entails significant ethical and institutional responsibilities: the sensitivity, scope, and potential impact of the data require careful reflection on data governance, privacy, fairness, and institutional accountability.

In our project, we develop a machine learning (ML)-based recommendation system described in Muehlbauer and Weber (2022), and Muehlbauer and Weber (2024) that seeks to support caseworkers by predicting both the *matching probability* - the likelihood that a given job seeker will be hired for a particular position - and the *job match quality*, i.e. job stability and wages. The innovation of our approach lies in the combination of both dimensions in generating recommendations, which aims to move beyond narrow notions of placement efficiency. By offering high-quality alternatives rather than focusing solely on the most probable matches, the system is designed to encourage jobseekers to explore a broader set of employment opportunities and to inform strategic counselling interventions.

The ethical and institutional implications of this system are substantial. In the following section (3.1), we first present the ethical strategy that underpins our approach, including the initial steps taken to implement it - an interview study with key stakeholders, which we introduce in detail in Section 3.1.1. In Chapter 4, we then turn to the concrete outcomes of our approach to ethical integration: we discuss the specific ethical dimensions currently under closer examination and how we are addressing them both analytically and technically - focusing in particular on data ethics and algorithmic bias, and the design of interpretable and accountable AI systems / XAI.

3.1 Ethical Strategy and First Implementation Steps in our Use Case

The introduction of algorithmic systems in the context of PES raises not only technical but also deeply social and ethical questions. As the international comparison in Chapter 2 has shown, such systems often fail due to concerns about discrimination - as in the cases of Poland and Austria - or due to a lack of acceptance by caseworkers, as seen in Switzerland, Finland, and South Korea. These experiences underscore that the success of AI in active labor market policy depends not only on predictive accuracy or technical robustness, but also on the capacity to anticipate and address normative tensions from the outset. In particular, the issues of non-discrimination and institutional trust must be foregrounded in both system design and implementation.

In our project, we address these challenges through a structured ethical framework grounded in the approach of *Embedded Ethics and Social Sciences* (EE), which has recently gained traction in

AI ethics research (McLennan et al., 2020, 2022). EE provides a framework for the ongoing integration of ethical and social reflection throughout the entire research and development process. Rather than positioning ethics as an external add-on or a post hoc evaluation, it emphasizes sustained collaboration between ethicists, social scientists, computer scientists, and domain experts. Within our team, this principle is already being operationalized by embedding ethical expertise directly into the project team - one of the authors of this paper serves as the designated ethicist working closely with technical and institutional stakeholders.

To further strengthen the integration of EE in our process, we draw on methodological guidance from the toolbox developed by Willem et al. (2024), which translates the abstract principles of EE into practical instruments for interdisciplinary teams. These tools enable the continuous reflection on values, risks, and social expectations at different stages of system development. They support the identification of ethically salient design choices and foster iterative, dialogue-oriented processes between technical developers and affected stakeholders.

However, while the EE framework already emphasizes the importance of integrating ethical reflection through stakeholder engagement—e.g., through qualitative interviews, workshops, and focus groups with those affected—it is particularly crucial in our use case to highlight this dimension. Acceptance by caseworkers and trust in the system are key conditions for successful implementation. Frontline professionals play a decisive role in interpreting, validating, and applying algorithmic outputs in their day-to-day practice. If their perspectives, concerns, and professional values are not taken seriously from the outset, there is a significant risk that they will resist the system or circumvent it in practice. (Dietz, forthcoming a; Marienfeldt, 2024).

Hence, involving these actors early and meaningfully is not just ethically desirable but strategically essential. The integration of stakeholder knowledge contributes to institutional trust and legitimacy, supports alignment with professional routines, and allows for a more nuanced understanding of the practical conditions under which the system will be used. This, in turn, enhances the interpretability and usability of the system's outputs. While issues of acceptance and trust are thus closely tied to stakeholder engagement, other normative concerns - particularly around non-discrimination - require additional attention to technical and conceptual dimensions. In Chapter 3.2, we examine in detail how we address these challenges, focusing on approaches related to bias mitigation, data ethics, and explainable AI (XAI).

3.1.1 Stakeholder Involvement through Qualitative Interviews

As a first step in operationalizing our EE-approach, we conducted a qualitative interview study with key stakeholders within the PES. This study aimed to elicit situated perspectives, normative expectations, and practical concerns surrounding the potential implementation of algorithmic decision-support systems in active labor market policy. In line with EE's emphasis on early and continuous engagement with affected actors, we focused in particular on the views of caseworkers, whose professional judgment and discretion are central to the functioning of the PES. Their insights are critical not only for understanding what ethical challenges may emerge in practice, but also for anticipating which factors influence trust, acceptance, and the perceived legitimacy of such systems. By foregrounding the lived expertise of practitioners, the study lays the groundwork for an iterative, reflexive development process that is attuned to both institutional logics and normative commitments.

The interview guide was developed in a collaborative process within our interdisciplinary team, bringing together technical expertise with ethical and social science perspectives. Our intention was to formulate questions that would not only capture the operational realities of system deployment but also open space for reflection on its normative dimensions. For the subsequent analysis, we drew on a category-based approach inspired by Mayring's method of qualitative content analysis (Mayring and Fenzl, 2022). We considered this framework particularly promising for practice-oriented research, as it provides both systematic guidance and sufficient flexibility. In applying this approach, our aim was to structure and reduce the material in a way that maintains its contextual richness while rendering it analytically tractable. The use of predefined thematic categories—further refined in the course of coding—was intended to facilitate the identification of relevant patterns without neglecting important nuances. In the following section, we do not elaborate on the coding scheme itself but rather point to empirical tendencies we consider especially significant for the ongoing development and ethically responsible implementation of the system.

As described earlier, Germany uses a caseworker-based profiling system (Desiere et. al., 2019), unlike some countries that rely on rule-based or statistical profiling. In some countries, there is also a mixed form of two or even all three profiling strategies. These different national approaches significantly influence not only the available kind of data, but also the design and implementation requirements of algorithmic systems. In Germany, for example, data protection regulations strictly limit the collection and use of certain sensitive variables for the placement process. Consequently, the model can only incorporate features that are legally permitted to be recorded, even if other variables would statistically improve performance.

The corresponding caseworkers interviewed work in different parts of the placement process like long-term or short-term unemployment, career advice, rehabilitation or employer service. The interviews contain three parts. First, we wanted to gain an understanding of the status quo and requirements in the current placement process. Second, we were interested in their openness to AI systems and if they considered them as useful. Then, we introduced the job recommendations provided by the system as described in chapter 3.0 and asked how they would assess them and whether this would be a useful aif in their work.

One of the key findings concerns the perception of the current matching process. The caseworkers describe the current matching system as too strict. They see space for improvements in terms of flexibility and completeness. A reduction in bureaucracy and routine tasks would also be desirable, allowing them to focus on their clients. The caseworkers describe various ways how they finally create their job recommendations. In general, they are open to AI technologies. The interviewed persons seemed not afraid of being replaced, because there are too many non-routine tasks and individual circumstances that have an influence on the recommendations. However, some believed there are colleagues with differing opinions. They consistently thought, that having algorithm-based job recommendations could have a positive effect on their work. In their opinion, it is important to provide appropriate introductions and information about the model to ensure acceptance also among colleagues with reservations. It is important to explain where the suggestions come from and what happens in the "black box". This can alleviate the concern that the algorithm provides arbitrary job recommendations.

The interview findings are consistent with those of Dietz et al. (forthcoming b). Caseworkers are more likely to accept AI-based recommendations if they are provided with detailed background information. In this context, this means providing information on the setup and how potentially discriminatory factors are handled. Caseworkers should be reassured that the intention is to provide support and not to replace them. The system's practical usefulness plays a key role in gaining acceptance.

The interview study evaluation is used to ensure that the system design can be optimally adapted to the needs of caseworkers. Concluding, they see potential for improving the current bidirectional matching system in terms of flexibility and efficiency. They are generally open to an algorithm-based matching tool that can speed up processes and achieve better results for jobseekers. In the best case, this would also allow caseworkers to focus more on the individual needs of jobseekers. Due to the caseworker-based profiling strategy in Germany, caseworkers can determine how intensively they use the system. Thus, the influence of the AI-system depends on the individual placement strategies, which can vary between caseworkers and depend on their personal preferences and experience.

To gauge reactions to algorithmic output, we presented participants with two example scenarios: fictional employment biographies and the corresponding top five job recommendations generated by the model. The interviewees generally regarded the higher-ranked suggestions as plausible and practical alternatives. Though regional adaptation was not included in the examples, they acknowledged the importance of incorporating local labor market data in real-world applications. Beyond standard job placements, they also identified potential for the system to support retraining decisions, vocational guidance, and career redirection. Importantly, they saw the combination of algorithmic input with personal assessment as a potential strength. By integrating insights from the model with their own professional judgment, they anticipated faster transitions into sustainable employment and a broader range of job alternatives for their clients. Compared to the current system - which they criticized for requiring extensive manual inputs and producing limited recommendations - the AI-based tool was seen as significantly more efficient and capable of capturing a fuller picture of the labor market.

As the interviews made clear, however, system acceptance does not stop with caseworkers. Jobseekers themselves must also be informed and empowered. There is a risk that algorithmic suggestions will be misunderstood either as binding or as irrelevant. Caseworkers emphasized the importance of clearly communicating that these are non-binding recommendations, which they will jointly assess with the jobseeker. Ensuring that the final decision remains with the individual not only aligns with legal and normative expectations but is also essential for trust. To ensure this, we base our design on shared decision-making in medical AI development. We will elaborate on this consideration in our specific implementation of XAI in 4.2.

4 Operationalizing Ethics – Social-Technical Design in Public Sector AI

The development and deployment of AI systems in the public sector, particularly within the domain of PES, raise several ethical challenges that can only be overcome through interdisciplinary cooperation. As discussed in Chapter 2, PES contexts are shaped by specific institutional, legal, and normative expectations, especially regarding fairness, accountability, and the protection of individual rights. In Germany, the use of AI-based systems in PES intersects with a particularly sensitive data environment: administrative labor market data that encompasses extensive information on jobseekers, including socio-demographic characteristics, employment histories, and benefit entitlements. This data volume and granularity create unique opportunities for predictive modelling, but they also impose a heightened responsibility concerning data ethics and governance. Importantly, the regulatory framework set by the EU AI Act must be fully observed. According to this legislation, AI systems are classified into four risk categories: unacceptable, high, limited, and minimal. Our case study falls under the category of high-risk AI, which is subject to a comprehensive set of obligations laid out in Articles 8 through 17. These include establishing a robust risk management framework, ensuring data governance, maintaining technical documentation, enabling logging and record-keeping, designing for human oversight, and providing sufficient levels of accuracy, robustness, and cybersecurity. Moreover, a quality management system must be in place to secure compliance with these requirements. Only if all of these conditions are met without exception can the system be considered for public sector deployment.

However, compliance with regulatory provisions alone does not guarantee ethical adequacy. While the EU AI Act sets necessary formal standards, it remains largely agnostic about the specific normative tensions that may arise in practice. For example, offering instructions for use that technically fulfil legal requirements does not automatically ensure that the system is comprehensible or trustworthy from the perspective of caseworkers or jobseekers. It is therefore essential to interpret these legal obligations within a broader ethical framework that addresses substantive concerns such as bias, opacity, and power asymmetries between users and automated systems. Of particular relevance for our use case is among others Article 56 of the EU AI Act, which explicitly states that AI systems in employment services are considered high-risk in part because they may ‘perpetuate historical patterns of discrimination, for example against women, certain age groups and persons with disabilities, or persons of a certain racial or ethnic origin or sexual orientation. (EU 2024, Recitals 56; 57) It must be critically noted that the phrasing ‘a certain racial origin’ is highly problematic. The very notion of ‘racial origin’ is conceptually flawed, as race is a social construct - comprising both external ascriptions and self-identification - that produces social realities. It must be clearly distinguished from, and not conflated with, biology or ancestry. However, the regulation nevertheless identifies key social dimensions where discrimination risks may materialize - namely, sexism, ageism, ableism, and racism. These categories point to structural injustices that may be reinforced through historical data and algorithmic reproduction of biased patterns.

This observation constitutes a core ethical concern of our work and serves as a transition to the subsequent section (4.1), where we explore in greater detail how our project addresses data ethics and the mitigation of algorithmic bias. Specifically, we analyze how discrimination risks can be understood and managed at the intersection of data practices and model design.

4.1 Data Ethics and Algorithmic Bias

The data used in German PES is derived from administrative records collected by caseworkers during in-person consultations at local job centers operated by the German Federal Employment Agency (BA). Job recommendations are currently based on a bidirectional matching system that aligns employer requirements with jobseeker profiles. This matching logic tends to be inflexible and restrictive: if employers specify too many requirements, or if jobseekers provide limited information, the system struggles to identify suitable matches. As a result, strict matching rules may inadvertently exclude a significant number of viable candidates or vacancies.

Naturally, this data contains discriminatory patterns because, as numerous scholars have shown, all societies - including Germany - are structured by intersecting systems of power such as racism, sexism, and ableism (Beigang et al., 2017; Crenshaw, 1991; Foroutan, 2021; Hill Collins, 2002). These structural inequalities are encoded in data and may be unintentionally reproduced - or even amplified - by algorithmic systems trained on such data.

To avoid the reinforcement of systematic disadvantages which is described as a form of statistical discrimination by the OECD (2019, p. 24), biases have to be identified and treated accordingly. To ensure that our use case also fulfils the criteria concerning data and data governance specified the EU AI Act, biases can be identified according to Ntoutsi et al. (2020). They distinguish between understanding, mitigating and accounting for bias. The understanding bias deals with the question how fairness has to be defined such that it can be considered in AI systems. Often, the data contains inequalities or discriminations that are reproduced by the algorithm. There could be sensitive features and their causal influences or over-representation of certain groups in the data. In this context, it is also an important to be sure how fairness is defined. The mitigating bias can address pre-processing (i.e. the data), in-processing (i.e. the ML algorithm) or post-processing (i.e. the ML model). There are different approaches that can be applied to mitigating bias. Accounting for bias can be either proactively or retroactively. There are many methods of bias-aware data collection to account for bias proactively and also to explain the AI decisions retroactively. Concluding, the authors describe a variety of technical challenges and solutions but also point out that biases are deeply embedded in our societies. Thus, the problem cannot be solved only with technical solutions.

To address the risk of perpetuating historical patterns of discrimination - explicitly identified in the EU AI Act (EU 2024, Recital 56) as a central concern in the context of high-risk systems - it is crucial to move beyond a narrow focus on formally protected attributes such as gender, race, age, or disability. While these categories rightly serve as central reference points in anti-discrimination efforts, a more comprehensive approach considers variables that may correlate with them and thereby serve as functional proxies within algorithmic models.

This issue becomes evident in the example of predictive policing systems in the United States. After concerns of racial discrimination were raised, the variable ‘race’ was removed from the

model as a technical remedy. However, discriminatory effects continued because other parameters - most notably postal codes - remained in use. These variables, though seemingly neutral, were strongly correlated with racial and socio-economic segregation, and thus continued to reproduce racial bias under a different guise. (Skeem and Lowenkamp, 2016) The case illustrates that eliminating protected characteristics from a dataset does not necessarily prevent discriminatory outcomes; it may instead obscure them and render their structural causes more difficult to detect. Such examples point to a broader issue. Attempting to neutralize algorithms by technically sanitizing input variables risks ignoring the social conditions in which discrimination occurs. Instead of concealing the existence of such structures, a more constructive approach identifies and examines them explicitly. In this sense, ethical algorithm design must go beyond mere compliance and strive for a deeper engagement with the forms of disadvantage it might reinforce or obscure.

In our project, we begin with a technical analysis to investigate possible discriminatory structures in the dataset and model behavior. Similar studies have been conducted in other national contexts - for example, a gender-focused analysis in France - which serve as useful starting points (Bied et al., 2023). These authors discuss gender fairness within the context of the audit of a recommender system called Multi-head Sparse E-recruitment. The study investigates whether gendered preferences for commuting time, contract type, wage, or other factors contribute to differences in recommendations for men and women. The main findings are that recall is slightly higher for women, and the authors provide evidence of differentiated treatment of men and women by the algorithm. However, these approaches often remain limited in scope. They tend to treat categories such as gender in isolation and thereby miss the more complex social constellations in which discrimination typically takes place.

To address these shortcomings, we adopt an intersectional perspective that is grounded in current research in the humanities and critical social sciences. As Kimberlé Crenshaw and others have shown, discriminatory experiences rarely arise from single social markers alone. Rather, they emerge at the intersection of multiple, overlapping dimensions - such as gender, race, class, disability and age - whose combined effects can reinforce marginalization in specific and often cumulative ways (Crenshaw, 1991). As such, an ethical and empirically adequate analysis of algorithmic bias must be capable of recognizing and analyzing these constellations.

In practical terms, this means that our assessment of fairness does not examine isolated variables, but considers subgroups defined by combinations of features. For instance, we will include profiles such as women over 50 of a particular nationality or persons with disabilities with interrupted employment histories. This allows us to identify discriminatory patterns that would remain invisible in a unidimensional framework and to respond with targeted, socio-technically informed design strategies. By following this approach, we aim to develop a system that meets not only the formal requirements of high-risk AI governance, but also the ethical obligations that arise from the real-world contexts in which such technologies are deployed.

While the specific implications of our intersectional bias analysis for socio-technical system design will be addressed in detail in a forthcoming publication, some critical points of discussion are already foreseeable and warrant preliminary consideration.

First, it is crucial to recognize that imbalances within the job placement process cannot be entirely eliminated. Certain occupations attract individuals with specific qualifications, life

experiences, or preferences that may correlate with gender or other social categories. Although gender, for instance, must be understood as a social construct - and although the model must avoid reproducing or reinforcing essentialist stereotypes - these social constructs nonetheless (can) shape the distribution of roles and choices in society. What we are confronted with here is not a simple causal relationship, but a complex interplay of self-selection, historical inequality, and normative expectation. Their influence must be critically examined and carefully constrained to ensure that they are not further entrenched through technological mediation. Moreover, during personal interviews, caseworkers get a lot of information that is not included in the data. This can have - more or less – a significant influence on the choice of job recommendations without being included in the statistical model. The objective must be to create systems that – together with caseworkers - remain open and responsive to a changing social reality.

This perspective also opens the door to a more constructive and reflexive role for algorithmic profiling in public systems. As the case of Sweden has shown, in some cases it may not only acceptable but even desirable for algorithmic systems to recognize patterns associated with specific social groups - especially those at higher risk of structural disadvantage. The conceptual utility of terms such as racism, sexism, or ableism lies precisely in their ability to expose the pervasive forms of discrimination and marginalization that often remain hidden in dominant narratives and institutional routines. Similarly, algorithmic profiling - when used critically and reflexively - can help expose these social inequalities by identifying statistical correlations that reflect real-world injustices. Rather than taking these patterns as grounds for exclusion, they can be incorporated into socially responsive design and targeted support within socio-technical infrastructures. In this sense, algorithmic systems can contribute to institutional reflection and reform, provided that their implementation takes place within a broader framework of ethical control, accountability, and participatory governance (Barocas et al., 2023a; D'Ignazio & Klein, 2020; Eubanks, 2017).

Against this background, it is worth reconsidering the common strategy of removing sensitive variables such as gender from algorithmic systems entirely. While this may initially seem like a straightforward anti-discrimination measure, it can undermine the system's practical relevance and effectiveness. Furthermore, as Bied et al. (2023) argue, algorithmic recommendations that deviate too far from jobseekers' actual search behaviors and preferences may lead to inefficiencies and, ultimately, to deadweight losses. If gender-specific search strategies are ignored, the system risks producing outcomes that are misaligned with real-world expectations and therefore less actionable for users. In such cases, a nuanced balance must be found - one that neither reinforces existing stereotypes nor ignores empirically significant patterns that could inform more inclusive and realistic placement strategies.

What can already be stated with confidence, however, is that regardless of the specific socio-technical design, the ethical inquiry surrounding this system does not - and must not - end with the development of a prototype, nor with its initial deployment phase. On the contrary, the EE approach we have adopted requires an ongoing, iterative engagement with the technology throughout its lifecycle. Ethical reflection, in this context, is not a preliminary step preceding implementation, but an integral and continuous part of the system's operational logic. Accordingly, the final design of the recommendation system will not be confined to the technical architecture of the algorithm itself. It will necessarily encompass institutional and procedural

components such as targeted training programmes for caseworkers, mentorship structures, continuous feedback loops, and the long-term institutional anchoring of responsible points of contact. These elements are essential to ensure that ethical oversight does not remain abstract, but is embedded in daily practice and responsive to concrete challenges as they arise. A detailed elaboration of these socio-technical governance measures will be provided in our following paper.

Before turning to that, however, the following chapter will address another key dimension of ethical system design - one that is closely intertwined with the concerns discussed thus far.

We now shift our focus to the domain of explainable artificial intelligence (XAI), which plays a central role not only in attempting to establish fairness, but also in facilitating meaningful human oversight and enabling trust among system users.

4.2 Fairness and Explainable AI (XAI)

There are several technical approaches aimed at fostering fairness in ML. Metrices like disparate impact, equalized odds or statistical parity difference are commonly used to evaluate how different demographic groups are affected by algorithmic decisions. Depending on the context, pre-processing techniques - such as reweighting datasets or generating synthetic data - can help mitigate imbalances prior to model training. Alternatively, fairness constraints can be embedded directly into the learning process (in-processing), for example by modifying the loss function or adjusting optimization procedures. Post-processing methods, which adjust model outputs after training, provide yet another avenue for intervention. The choice between pre-, in-, or post-processing - or a combination of these - depends on the characteristics of the dataset, model architecture, and the normative framing of the research question.

However, while such technical interventions are essential, they are not sufficient on their own. As emphasized in the previous chapter, a purely technical perspective risks overlooking the broader social, institutional, and normative dimensions of fairness. Achieving fairness in AI systems - particularly in sensitive public sector contexts - requires interdisciplinary reflection and dialogue, combining insights from computer science, social theory, and ethics. This broader understanding of fairness also has implications for how algorithmic systems are communicated and understood in practice. In this context, explainability becomes not only a technical challenge but a prerequisite for meaningful accountability, trust, and informed human oversight. This section therefore turns to the role of XAI within our project, outlining both its practical relevance for end users and the conceptual tensions involved in making complex models transparent and interpretable.

By generating interpretable representations of model behavior, XAI could help uncover biases embedded in the data or the model's decision logic. However, while promising, these methods are far from uncontroversial. XAI in particular has been criticized for being positioned as a cure-all solution, even though it cannot, by itself, guarantee fairness or prevent discrimination (Deck et al., 2024; Ghassemi et al., 2021; Schemmer et al., 2023).

Freiesleben and König (2023) show some critical aspects in current XAI research. They describe some key misconceptions like: explanation methods are purpose-free, there is one explanation technique to rule them all, benchmarks do not need a ground-truth, people should get

explanations they find intuitive, current deep nets accidentally learn human concepts, every XAI paper needs human studies, XAI methods can be wrong and extrapolating to stay true to the model. They recommend notapplying a standardized approach, but rather thinking carefully about which approach is the right one.

In designing our approach to XAI, we draw on developments in the field of medical AI - particularly those associated with *shared decision-making*. As in medicine, our objective is not to reduce, but to strengthen the agency of those affected by algorithmic recommendations. If explainability is directed solely at caseworkers, this risks reinforcing existing asymmetries in the counselling process. A jobseeker confronted with a recommendation issued by a counsellor who is supported by a system they themselves do not understand is placed in a highly asymmetrical position - one that undermines the idea of informed decision-making.

By contrast, if explainability is designed in a way that enables counsellors and jobseekers to reflect jointly on a recommendation („*Does this suggestion make sense to you?*“) the interaction becomes more dialogical. It opens up space for critical assessment and contextualization, and enables jobseekers to engage with the suggestions on their own terms. The goal is not to make the algorithm appear neutral or authoritative, but to make its logic understandable and contestable. This requires a shift from one-sided to shared interpretability. Accordingly, in our use case, developing explainability must involve both user groups - counsellors and jobseekers - and aim for intuitive accessibility for both. Only under these conditions can explainability contribute meaningfully to fairness, trust, and human oversight.

Importantly, Freiesleben and König's list of critical aspects of XAI also calls for caution against the potential of XAI to obscure rather than clarify fairness issues. For instance, it is possible to generate user-facing explanations that omit sensitive variables such as gender, even when they heavily influence the underlying model's predictions. In such cases, the explanation may create a false impression of neutrality, thereby concealing discriminatory patterns embedded in the model. Rather than promoting accountability, XAI can thus inadvertently legitimize biased outcomes - unless its design and application are critically aligned with ethical and fairness considerations.

Weber et al. (2024) describe four challenges related to XAI, namely, disagreements on the scope of XAI, the lack of definitional cohesion, precision, and adoption, the issues with motivations for XAI research, and limited and inconsistent evaluations. These problems can be eliminated by taking measures. The authors recommend that researchers of different disciplines also consider the perspective of researchers with different conceptualizations. Zerilli (2022) contributes a valuable philosophical account of XAI that complements and deepens this critique. He argues that the opacity of modern machine learning systems has led to a proliferation of explanatory demands, which XAI research seeks to meet by balancing three competing desiderata: completeness (i.e., exhaustive or sufficiently deep representations of how a system works), realism or fidelity (faithfulness to the system's actual decision-making processes), and interpretability (the capacity of an explanation to be understood and used by affected individuals). His analysis underscores that the goal of XAI is not merely to offer technically accurate accounts of algorithmic processes, but to justify automated decisions in ways that are meaningful and actionable for the decision subjects themselves - especially in contexts where the right to contest or appeal decisions is at stake.

Similar critiques apply to the other technical strategies mentioned above. Weinberg (2022) discusses a variety of critiques of current fairness-enhancing technical interventions in ML. She also proposes solutions like incorporating causal graphs or intersectionality into technical fairness measures, redirecting the problem formulation space, the implementation of fairness checklists, improving data collection practices, creating protective optimization technologies, addressing the social context and the interdisciplinarity or cross-disciplinary collaboration for new methodologies. Barocas et al. (2023b) find that traditional tests for discrimination in combination with fairness studies of various algorithmic systems are well-suited for a single decision system at a single point of time. For more complex questions, a broader view of fairness aspects is necessary.

For these reasons, technical adjustments alone are insufficient to ensure fairness - a point we already emphasized in the previous chapter on bias mitigation, and which applies equally to the development of XAI. Neither of these components can, in themselves, guarantee fairness. However, both are essential prerequisites for achieving it - alongside trust, human oversight, and broader social accountability. What is also required, though, is a more comprehensive understanding of the socio-technical environment in which AI systems are embedded and operate. Decisions about how to implement XAI and fairness-oriented measures in specific models cannot be made in the abstract; they must be grounded in the specific social contexts in which these systems are deployed. One of the central insights here is that AI systems do not discriminate in isolation - and should not be evaluated as if they did. Discrimination arises within the broader institutional and social configurations in which these technologies are developed, applied, and interpreted. If we are serious about reducing discriminatory outcomes, we must interrogate and actively shape these surrounding contexts.

4.3 Implications for other Public Sector AI Systems

While the case study presented in this paper is situated within the German PES, many of the findings and recommendations from this paper's use case can be transferred to other contexts with certain adaptations. With appropriate contextual adjustments, the ethical and technical principles outlined here can be adapted for use in other national PES systems, as well as in different areas of public service provision. For example, AI can improve the effectiveness of policy making, the selection of suitable candidates for vocational training or efficiency of internal operations.

Labor market-specific characteristics must be adapted to the specifications applicable to the relevant social area and the available data. Thus, our findings on data preparation and the design of algorithms and output, as well as the insights into how the AI recommendations are introduced to the target group, can be relevant for many use cases in the PES and the social sector in general.

Although institutional structures, regulatory environments, and sociotechnical configurations differ between countries, the fundamental challenges associated with fairness, accountability, and human oversight in public sector AI systems are widely shared. These include, for instance, issues of data sensitivity, asymmetries of power between users and institutions, and the risks of reinforcing structural discrimination. As such, the core of our approach - embedding ethical

considerations throughout the development and deployment process - offers a transferable framework that can inform similar efforts elsewhere.

In particular, our recommendations regarding the development of discrimination-aware algorithmic models, and the design of transparent and participatory interfaces for delivering AI-based recommendations may have broader relevance. These elements can guide how public institutions approach AI integration in ways that are not only technically robust but also socially legitimate and ethically sound.

Naturally, the public sector is highly diverse, giving rise to a wide range of potential AI applications. This includes adaptations in model architecture, estimation techniques, institutional implementation processes, and legal frameworks. Still, a key conclusion of this study is that all such models, regardless of the sector, ultimately intervene in human lives. They are, by nature, high-stakes systems. For this reason, the imperative to avoid discrimination, ensure meaningful human oversight, and foster public trust should be viewed as foundational across all public sector AI initiatives. By grounding AI development in a situated understanding of local practices and user perspectives, and by drawing on interdisciplinary methodologies as demonstrated in this study, public sector actors can work toward building socio-technical systems that not only enhance administrative efficiency but also uphold democratic values and social equity.

5 Conclusion and Outlook

AI holds considerable potential to support many areas of the public sector, particularly within PES. However, successful implementation requires careful attention to both technical and ethical considerations. In this paper, we illustrate how AI can be designed to integrate social and ethical perspectives from the outset, using the Embedded Ethics and Social Sciences approach. This method emphasizes the inclusion of practitioners and stakeholders during system development, ensuring that socio-technical realities, user needs, and regulatory requirements are accounted for.

Our German PES use case demonstrates that ethically-informed AI design has the potential to improve job matching, provide broader options for jobseekers, and address potential discrimination without reinforcing existing biases. Our contribution to the existing literature is to show how AI systems must be designed to be successfully introduced in the public sector. The literature review indicates that there are various challenges in the pre-, in-, and post-estimation phase. It is crucial to adapt the model carefully to directives such as the EU AI Act and other specific requirements. This mainly concerns the technical design of the model. The general design and the introduction of the AI system are key points for achieving acceptance of the model. The design must strike a balance between providing an algorithm that supports employees in the respective public sector and avoiding overly restrictive interventions that could instil a fear of replacement. The willingness of users, along with sufficient background information, is crucial for ensuring the acceptance of the group of people affected by an AI system's outcome. Key socio-technical elements include the consideration of fairness and data ethics, transparent and explainable AI, and participatory practices that strengthen human oversight and user trust. By embedding ethical reflection throughout development, the system fosters acceptance among caseworkers and promotes agency for the individuals affected.

Despite these contributions, several limitations must be acknowledged. Controlling for discrimination and fairness is not straightforward. Our analysis shows that public sector AI is multifaceted. The analysis primarily draws on initial qualitative insights and conceptual frameworks, and the technical implementation of fairness measures and XAI is still in progress. Finally, the long-term effects on organizational workflows, user behavior, and actual outcomes for jobseekers remain to be evaluated empirically.

Looking forward, our next paper will focus on the **technical implementation of these principles**, including how intersectional fairness analysis is applied and how XAI is operationalized in practice. We will describe how the algorithm is adapted to the PES context and outline evaluation plans to assess system performance, user acceptance, and the effectiveness of the socio-technical design.

References

Allhutter, D., Alushi, A., Cavalcanti De Alcântara, R., Männiste, M., Pentzold, C., & Sosnowski, S. (2024). Public value in the making of automated and datafied welfare futures. Internet Policy Review, 13(3). https://doi.org/10.14763/2024.3.1803

Alon-Barkat, S., & Busuioc, M. (2023). Human–AI Interactions in Public Sector Decision Making: "Automation Bias" and "Selective Adherence" to Algorithmic Advice. Journal of Public Administration Research and Theory, 33(1), 153–169. https://doi.org/10.1093/jopart/muac007

Arni, P., & Schiprowski, A. (2016). Die Rolle von Erwartungshaltungen in der Stellensuche und der RAV-Beratung : Teilprojekt 2: Pilotprojekt Jobchancen-Barometer: Erwartungshaltungen der Personalberatenden, Prognosen der Arbeitslosendauern und deren Auswirkungen auf die Beratungspraxis und den Erfolg der Stellensuche. Staatssekretariat für Wirtschaft SECO. https://doi.org/10.21256/zhaw-30297

Barocas, S., Hardt, M., & Narayanan, A. (2023a). Fairness and machine learning. MIT Press. https://fairmlbook.org/

Barocas, S., Hardt, M., & Narayanan, A. (2023b). Fairness and machine learning: Limitations and opportunities. The MIT Press.

Beigang, S., Fetz, K., Kalkum, D., & Otto, M. (2017). Diskriminierungserfahrungen in Deutschland. Ergebnisse einer Repräsentativ- und einer Betroffenenbefragung. Antidiskriminierungsstelle des Bundes.

Bied, G., Gaillac, C., Hoffmann, M., Caillou, P., Crépon, B., Nathan, S., & Sebag, M. (2023). Fairness in job recommendations: Estimating, explaining, and reducing gender gaps. 3523. https://inria.hal.science/hal-04438512

Broecke, S. (2023). Artificial intelligence and labour market matching (OECD Social, Employment and Migration Working Papers 284). https://doi.org/10.1787/2b440821-en

Bullock, J. B. (2019). Artificial Intelligence, Discretion, and Bureaucracy. The American Review of Public Administration, 49(7), 751–761. https://doi.org/10.1177/0275074019856123

Caswell, D., Marston, G., & Larsen, J. E. (2010). Unemployed citizen or 'at risk' client? Classification systems and employment services in Denmark and Australia. Critical Social Policy, 30(3), 384–404. https://doi.org/10.1177/0261018310367674

Crenshaw, K. (1991). Mapping the Margins: Intersectionality, Identity Politics, and Violence against Women of Color. Stanford Law Review, 43(6), 1241. https://doi.org/10.2307/1229039

Dasgupta, R., Mekala, S. H., Jaigirdar, F. T., Anwar, A., & Chang, L. Y.-C. (2025). Unlocking Australia's AI usage in law enforcement from human involvement perspective: A systematic literature review. AI & SOCIETY. https://doi.org/10.1007/s00146-025-02350-6

Deck, L., Schoeffer, J., De-Arteaga, M., & Kühl, N. (2024). A Critical Survey on Fairness Benefits of Explainable AI. The 2024 ACM Conference on Fairness, Accountability, and Transparency, 1579–1595. https://doi.org/10.1145/3630106.3658990

Desiere, S., Langenbucher, K., & Struyven, L. (2019). Statistical profiling in public employment services: An international comparison (OECD Social, Employment and Migration Working Papers 224). https://doi.org/10.1787/b5e5f16e-en

Desiere, S., & Struyven, L. (2021). Using Artificial Intelligence to classify Jobseekers: The Accuracy-Equity Trade-off. Journal of Social Policy, 50(2), 367–385. https://doi.org/10.1017/S0047279420000203

Deutscher Ethikrat (Hrsg.). (2023). Mensch und Maschine—Herausforderungen durch Künstliche Intelligenz. Deutscher Ethikrat.

Dietz, Martin; Osiander, Christopher; Sirman-Winkler, Mareike; Tepe, Markus (forthcoming a): Biased by Design? Case Managers' Multidimensional Preferences Toward the Design of Algorithmic Decision Support Systems.

Dietz, Martin; Osiander, Christopher; Sirman-Winkler, Mareike; Tepe, Markus (forthcoming b): Are Case Managers Willing to Entrust Counseling Tasks to Artificial Intelligence? The Role of Administrative Discretion, Client Interaction, and Technology Attitudes.

D'Ignazio, C., & Klein, L. F. (2020). Data feminism. The MIT Press.

Dolata, M., Schenk, B., Fuhrer, J., Marti, A., & Schwabe, G. (2020). When the System Does Not Fit: Coping Strategies of Employment Consultants. Computer Supported Cooperative Work (CSCW), 29(6), 657–696. https://doi.org/10.1007/s10606-020-09377-x

Eubanks, V. (2017). Automating inequality: How high-tech tools profile, police, and punish the poor (First Edition). St. Martin's Press.

European Commission. (o. J.-a). European Network of Public Employment Services [Official website of the European Union]. Employment, Social Affairs and Inclusion. Abgerufen 27. Mai 2025, von https://employment-social-affairs.ec.europa.eu/policies-and-activities/coordination-employment-and-social-policies/european-network-public-employment-services_en

European Commission. (o. J.-b). The European Pillar of Social Rights in 20 principles. Employment, Social Affairs and Inclusion. Abgerufen 27. Mai 2025, von https://employment-social-affairs.ec.europa.eu/european-pillar-social-rights-20-principles_en

Europäisches Parlament und Europäischer Rat (13.06. 2024). Verordnung (EU) 2024/1689 zur Festlegung harmonisierter Vorschriften für künstliche Intelligenz und zur Änderung der Verordnungen (EG) Nr. 300/2008, (EU) Nr. 167/2013, (EU) Nr. 168/2013, (EU) 2018/858, (EU), 2018/1139 und (EU) 2019/2144 sowie der Richtlinien 2014/90/EU, (EU) 2016/797 und (EU) 2020/1828 (Verordnung über künstliche Intelligenz). Deutsche Fassung. http://data.europa.eu/eli/reg/2024/1689/oj

Fisher, E., O'Rourke, M., Evans, R., Kennedy, E. B., Gorman, M. E., & Seager, T. P. (2015). Mapping the integrative field: Taking stock of socio-technical collaborations. Journal of Responsible Innovation, 2(1), 39–61. https://doi.org/10.1080/23299460.2014.1001671

Fisher, E., & Schuurbiers, D. (2013). Socio-technical Integration Research: Collaborative Inquiry at the Midstream of Research and Development. In N. Doorn, D. Schuurbiers, I. Van De Poel, & M. E. Gorman (Hrsg.), Early engagement and new technologies: Opening up the laboratory (Bd. 16, S. 97–110). Springer Netherlands. https://doi.org/10.1007/978-94-007-7844-3_5

Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., Luetge, C., Madelin, R., Pagallo, U., Rossi, F., Schafer, B., Valcke, P., & Vayena, E. (2018). AI4People—An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations. Minds and Machines, 28(4), 689–707. https://doi.org/10.1007/s11023-018-9482-5

Foroutan, N. (2021). Die postmigrantische Gesellschaft: Ein Versprechen der pluralen Demokratie (2. Aufl.). transcript Verlag. https://doi.org/10.14361/9783839459447

Freier, C., & Senghaas, M. (2022). Placement Advisors as Innovators. How Professionals Use Enhanced Discretion in Germany's Public Employment Services. Journal of Social Policy, 51(1), 155–172. https://doi.org/10.1017/S0047279420000744

Freiesleben, T., & König, G. (2023, Juni 7). Dear XAI Community, We Need to Talk! Fundamental Misconceptions in Current XAI Research. arXiv.Org. https://arxiv.org/abs/2306.04292v1

Friedman, B., Kahn, P., & Borning, A. (2002). Value Sensitive Design: Theory and Methods | PDF. Scribd. https://www.scribd.com/document/448586629/Value-Sensitive-Design-Theory-and-Methods

Gesk, T. S., & Leyer, M. (2022). Artificial intelligence in public services: When and why citizens accept its usage. Government Information Quarterly, 39(3), 101704. https://doi.org/10.1016/j.giq.2022.101704

Ghassemi, M., Oakden-Rayner, L., & Beam, A. L. (2021). The false hope of current approaches to explainable artificial intelligence in health care. The Lancet Digital Health, 3(11), e745–e750. https://doi.org/10.1016/S2589-7500(21)00208-9

Gottwald, M., & Sowa, F. (2019). 'A little more humanity': Placement officers in Germany between social work and social policy. In M. Gottwald & F. Sowa, Social Work and the Making of Social Policy (S. 201–216). Policy Press. https://doi.org/10.1332/policypress/9781447349150.003.0013

Gräfe, P., Marienfeldt, J., Wehmeier, L. M., & Kuhlmann, S. (2024). Changing tasks and changing public servants? The digitalisation and automation of public administrative work. Information Polity, 13837605241289773. https://doi.org/10.1177/13837605241289773

Hartl, T., Hutter, C., & Weber, E. (2021). Matching for three: Big data evidence on search activity of workers, firms, and employment service. IAB Discussion Paper.

Hill Collins, P. (2002). Black Feminist Thought (0 Aufl.). Routledge. https://doi.org/10.4324/9780203900055

Jörg, S., Ziethmann, P., & Breuer, S. (2023). MedAIcine: A Pilot Project on the Social and Ethical Aspects of AI in Medical Imaging. HCI International 2023 Posters. HCII2023.

Kolleck, A., & Orwat, C. (2020). Mögliche Diskriminierung durch algorithmische Entscheidungssysteme und maschinelles Lernen – ein Überblick [PDF]. Büro für Technikfolgen-Abschätzung beim Deutschen Bundestag (TAB). https://doi.org/10.5445/IR/1000127166

Körtner, J., & Bonoli, G. (2021). Predictive Algorithms in the Delivery of Public Employment Services. https://doi.org/10.31235/osf.io/j7r8y

Marienfeldt, J. (2024). Does Digital Government Hollow out the Essence of Street-level Bureaucracy? A Systematic Literature Review of How Digital Tools' Foster Curtailment, Enablement and Continuation of Street-level Decision-making. In Social Policy & Administration 58, no. 5 (2024): 831–855. https://onlinelibrary.wiley.com/doi/10.1111/spol.12991

Masmoudi, M., Jarboui, B., & Siarry, P. (Hrsg.). (2021). Artificial Intelligence and Data Mining in Healthcare. Springer International Publishing. https://doi.org/10.1007/978-3-030-45240-7

Mayring, P., & Fenzl, T. (2022). Qualitative Inhaltsanalyse. In N. Baur & J. Blasius (Hrsg.), Handbuch Methoden der empirischen Sozialforschung (S. 691–706). Springer Fachmedien Wiesbaden. https://doi.org/10.1007/978-3-658-37985-8_43

McLennan, S., Fiske, A., Celi, L. A., Müller, R., Harder, J., Ritt, K., Haddadin, S., & Buyx, A. (2020). An embedded ethics approach for AI development. Nature Machine Intelligence, 2(9), 488–490. https://doi.org/10.1038/s42256-020-0214-1

McLennan, S., Fiske, A., Tigard, D., Müller, R., Haddadin, S., & Buyx, A. (2022). Embedded ethics: A proposal for integrating ethics into the development of medical AI. BMC Medical Ethics, 23(1), 6. https://doi.org/10.1186/s12910-022-00746-3

Minocher, X., & Randall, C. (2020). Predictable policing: New technology, old bias, and future resistance in big data surveillance. Convergence: The International Journal of Research into New Media Technologies, 26(5–6), 1108–1124. https://doi.org/10.1177/1354856520933838

Mühlbauer, S., & Weber, E. (2022). Machine Learning for Labour Market Matching. IAB-Discussion Paper, 202203. https://doi.org/10.48720/IAB.DP.2203

Mühlbauer, S., & Weber, E. (2024). Predicting Job Match Quality: A Machine Learning Approach. IAB-Discussion Paper, 202409. https://doi.org/10.48720/IAB.DP.2409

Ntoutsi, E., Fafalios, P., Gadiraju, U., Iosifidis, V., Nejdl, W., Vidal, M., Ruggieri, S., Turini, F., Papadopoulos, S., Krasanakis, E., Kompatsiaris, I., Kinder-Kurlanda, K., Wagner, C., Karimi, F., Fernandez, M., Alani, H., Berendt, B., Kruegel, T., Heinze, C., … Staab, S. (2020). Bias in data-driven artificial intelligence systems—An introductory survey. WIREs Data Mining and Knowledge Discovery, 10(3), e1356. https://doi.org/10.1002/widm.1356

Osiander, C., & Steinke, J. (2011). Street-level bureaucrats in der Arbeitsverwaltung: Dienstleistungsprozesse und reformierte Arbeitsvermittlung aus Sicht der Vermittler. Zeitschrift Für Sozialreform, 57(2), 149–173. https://doi.org/10.1515/zsr-2011-0203

Schemmer, M., Kuehl, N., Benz, C., Bartos, A., & Satzger, G. (2023). Appropriate Reliance on AI Advice: Conceptualization and the Effect of Explanations. Proceedings of the 28th International Conference on Intelligent User Interfaces, 410–422. https://doi.org/10.1145/3581641.3584066

Schlögl-Flierl, K., & Ziethmann, P. (2025). KI und Wir: Warum der Einsatz von Künstlicher Intelligenz das Prinzip der Multiakteursverantwortung braucht und forciert. In S. Böhmer & T. Unger (Hrsg.), Technisches Handeln und Verantwortung (S. 35–53). Springer Fachmedien Wiesbaden. https://doi.org/10.1007/978-3-658-46631-2_3

Scupola, A., & Mergel, I. (2022). Co-production in digital transformation of public administration and public value creation: The case of Denmark. Government Information Quarterly, 39(1), 101650. https://doi.org/10.1016/j.giq.2021.101650

Selten, F., & Klievink, B. (2024). Organizing public sector AI adoption: Navigating between separation and integration. Government Information Quarterly, 41(1), 101885. https://doi.org/10.1016/j.giq.2023.101885

Skeem, J. L., & Lowenkamp, C. T. (2016). RISK, RACE, AND RECIDIVISM: PREDICTIVE BIAS AND DISPARATE IMPACT*. Criminology, 54(4), 680–712. https://doi.org/10.1111/1745-9125.12123

Van Der Burg, S. (2009). Imagining the Future of Photoacoustic Mammography. Science and Engineering Ethics, 15(1), 97–110. https://doi.org/10.1007/s11948-008-9079-0

Weber, R. O., Johs, A. J., Goel, P., & Silva, J. M. (2024). XAI is in trouble. AI Magazine, 45(3), 300–316. https://doi.org/10.1002/aaai.12184

Weinberg, L. (2022). Rethinking Fairness: An Interdisciplinary Survey of Critiques of Hegemonic ML Fairness Approaches. Journal of Artificial Intelligence Research, 74, 75–109. https://doi.org/10.1613/jair.1.13196

Wenzelburger, G., König, P. D., Felfeli, J., & Achtziger, A. (2024). Algorithms in the public sector. Why context matters. Public Administration, 102(1), 40–60. https://doi.org/10.1111/padm.12901

Willem, T., Fritzsche, M.-C., Zimmermann, B. M., Sierawska, A., Breuer, S., Braun, M., Ruess, A. K., Bak, M., Schönweitz, F. B., Meier, L. J., Fiske, A., Tigard, D., Müller, R., McLennan, S., & Buyx, A. (2024). Embedded Ethics in Practice: A Toolbox for Integrating the Analysis of Ethical and Social Issues into Healthcare AI Research. Science and Engineering Ethics, 31(1), 3. https://doi.org/10.1007/s11948-024-00523-y

Williams, B. A., Brooks, C. F., & Shmargad, Y. (2018). How Algorithms Discriminate Based on Data They Lack: Challenges, Solutions, and Policy Implications. Journal of Information Policy, 8, 78–115. https://doi.org/10.5325/jinfopoli.8.2018.0078

Yollu-Tok, A., Beblo, M., Draude, C., Gegenhuber, T., Höyng, S., Nebe, K., Send, H., Spiecker, I., & Teubner, T. (2021). Digitalisierung geschlechtergerecht gestalten—Dritter Gleichstellungsbericht. Bundesministerium für Familie, Senioren, Frauen und Jugend. https://www.bmfsfj.de/blob/jump/184544/dritter-gleichstellungsbericht-bundestagsdrucksache-data.pdf

Zerilli, J. (2022). Explaining Machine Learning Decisions. Philosophy of Science, 89(1), 1–19. https://doi.org/10.1017/psa.2021.13

Ziethmann, P., Elia, M., Stieler, F., Bauer, B., Welzel, J., & Schlögl-Flierl, K. (2025). Clinical Decision Support Systems at the Intersection of Technology and Ethics: A Critical Analysis of the Ethical Guidelines Issued by the German Medical Association. Digital Society, 4(1), 15. https://doi.org/10.1007/s44206-025-00175-w

Imprint

IAB-Discussion Paper 12|2025

Date of publication
October 1, 2025

Publisher
Institute for Employment Research
of the Federal Employment Agency
Regensburger Str. 104
90478 Nürnberg Germany

Download of this IAB-Discussion Paper
https://doku.iab.de/discussionpapers/2025/dp1225.pdf

All publications in the series "IAB-Discussion Paper" can be downloaded from
https://iab.de/en/publications/iab-publications/iab-discussion-paper-en/

Website
https://iab.de/en/

ISSN
2195-2663

DOI
10.48720/IAB.DP.2512

Corresponding author
Sabrina Mühlbauer
Phone: +49 911 179-9743
Email: sabrina.muehlbauer@iab.de
Paula Ziethmann
Phone: +49 911 179-4922
Email: paula.ziethmann@iab.de