dortmund university

Responsible AI in Human Resource Management

An in-depth case examination of transparent and fair use of Machine Learning

DISSERTATION

by

Ansgar Heidemann

DISSERTATION

to obtain the academic degree Doctor rerum politicarum (Dr. rer. pol.)

submitted at the

Faculty of Business and Economics

1. Reviewer: Prof. Dr. Andreas Hoffjan
2. Reviewer: Prof. Dr. Andreas Liening
3. Member of the commission: Prof. Dr. Jens Rowold

Dortmund

2025

Acknowledgements

Writing this dissertation has been a long and challenging five-year journey, and I could not have done it without the incredible support of so many people. First and foremost, I want to sincerely thank **Prof. Dr. Andreas Hoffjan** for his outstanding guidance, mentorship, and exceptional ability to bring together the right opportunities, topics, and people so they can fully realize their potential. His advice and encouragement have been invaluable, and I truly appreciate the time and effort he has invested in helping me navigate this process. I am equally incredibly grateful to my co-authors, **Dr. Michael Tekieli**, **Christian Ertel**, and especially **Svenja Hülter**, for the productive collaboration. Working together has been an absolute pleasure - our teamwork in which everyone was able to contribute their strengths not only led to great results but also made the journey much more enjoyable.

A huge thank you to all my colleagues at **Windhoff Group**, whose support in resources and networks has made a real difference along the way. I want to give a special shout-out to my supervisors, **Matthias Kappelhoff**, **David Baumeister**, and **Michael Hornhues**, for believing in me and giving me the opportunity to take on new challenges. Their trust and encouragement have meant the world to me. To my **family**, I cannot thank you enough for always being there for me - not only during my doctorate, but also throughout my 10 years at the university. And last, but the most important—my wife, **Joke**. Words can not fully express how grateful I am for your love, support, and understanding how much time I have invested in this project. You have been my anchor throughout this journey and I could not have done it without you.

Thank you all!

Dortmund

2025

Contents

Figures

Tables

Abbreviations

AI	Artificial Intelligence
ALE	Accumulated Local Effects
AF	Algorithmic Fairness
HRA	Human Resource Analytics
EU AI Act	European Union Artificial Intelligence Act
HRIS	Human Resource Information System
HRM	Human Resource Management
IS	Information Systems
ML	Machine Learning
PBC	Perceived Behavioural Control
PFI	Perturbated Feature Importance
SHAP	SHapley Additive exPlanations
XAI	Explainable Artificial Intelligence

Summary

This thesis explores the opportunities and risks of Machine Learning (ML) and Artificial Intelligence (AI) technology applied to high-risk decision-making and how technological and business measures contribute to responsible and value-adding implementation. The three included studies cumulatively examine a case study in a German federal agency that uses ML to predict voluntary employee turnover. This practice-oriented approach provides new empirical insights for understanding the successful design and implementation of ML-based algorithmic HRM from the perspectives of (1) transparency, (2) diversity, non-discrimination & fairness, and (3) individual adoption, with a separate study dedicated to each perspective. A key finding of the thesis is that technical measures such as explainable AI actually contribute to meeting the central requirements for the responsible use of AI. However, these technical measures are only effective if they are successfully implemented in the business organisation and processes under trained and critically thinking human supervision, thus outlining the importance and consequences of the European Union's current regulatory efforts. Thereby, this thesis contributes to our understanding of a highly topical issue, as the example of ChatGPT clearly shows that the world is changing at a rapid pace, with the overarching goal of leading to an AI generation that benefits individuals, organisations and society alike.

Zusammenfassung (German)

Diese Dissertation untersucht die Chancen und Risiken der Technologien „Maschinelles Lernen“ und der „Künstliche Intelligenz“ (KI) bei risikoreichen Entscheidungen und wie sowohl technologische als auch unternehmerische Maßnahmen zu einer verantwortungsvollen und wertschöpfenden Einsatz dieser Technologien beitragen. Die drei einbezogenen Studien untersuchen kumulativ eine Fallstudie in einer deutschen Bundesbehörde, die ML zur Vorhersage der freiwilligen Mitarbeiterfluktuation einsetzt. Dieser praxisorientierte Ansatz liefert neue empirische Erkenntnisse für das Verständnis der erfolgreichen Gestaltung und Implementierung von ML-basiertem algorithmischen HRM aus den Perspektiven (1) Transparenz, (2) Diversität, Nicht-Diskriminierung & Fairness und (3) individuelle Nutzung, wobei jeder Perspektive eine eigene Studie gewidmet ist. Ein zentrales Ergebnis der Arbeit ist, dass technische Maßnahmen wie erklärbare KI tatsächlich dazu beitragen, die zentralen Anforderungen an den verantwortungsvollen Einsatz von KI zu erfüllen. Diese technischen Maßnahmen sind jedoch nur dann wirksam, wenn sie in der Organisation und in den Prozessen unter geschulter und kritisch denkender menschlicher Aufsicht erfolgreich umgesetzt werden, was die Bedeutung und die Folgen der aktuellen Regulierungsbemühungen der Europäischen Union verdeutlicht. Damit trägt diese Studie zu unserem Verständnis eines hochaktuellen Themas bei, denn das Beispiel ChatGPT zeigt deutlich, dass sich die Welt in rasantem Tempo verändert, mit dem übergeordneten Ziel, eine KI-Generation zu erreichen, von der Individuen, Organisationen und die Gesellschaft gleichermaßen profitieren.

1 Introduction

1.1 Motivation

Since 2022, we have witnessed unprecedented growth in the possibilities and reach of AI – a revolutionary and disruptive technology that offers significant opportunities and challenges for organisations, society and individuals. One remarkable example of this burgeoning AI landscape is the rapid proliferation and hype surrounding Chat Generative Pre-trained Transformers (Chat GPT). This chat bot, empowered by ML, has become emblematic of AI's capabilities, and it allows machines to comprehend and generate human-like text with fluency and coherence in various domains (Brown et al. 2020; Dwivedi et al. 2023). The knowledge embedded in foundational large language models, built using ML algorithms such as neural network-based deep learning, has been proven capable of passing the final exams for higher education degrees in business management (Terwiesch 2023), law (Choi et al. 2023) and medicine (Gilson et al. 2023). The software, released on November 30 in 2022, is the fastest internet app to reach 100 million monthly active users, which it achieved within two months. The success of Chat GPT illustrates the enormous speed at which AI is evolving and spreading, surpassing even the previous records set by websites such as the social networks Instagram and TikTok (Time 2023).

The upside of AI manifests in the transformation of individual lives, organisations and the global impact. On an individual level, generative AI such as ChatGPT increases productivity in terms of efficiency by an average 14% to 56% (depending on task characteristics and prior knowledge), as well as the quality of results by about 20%, while also increasing enjoyment in the fulfilment of tasks (Noy and Zhang 2023; Peng et al. 2023; Brynjolfsson et al. 2023). Similar to the Industrial Revolution in the 1800s, when machines and robots began to automate manual tasks by going beyond the physical capabilities of humans, AI offers transformative potential to complement and eventually replace human intellectual and social applications (Dwivedi et al. 2021). The affected tasks particularly involve the jobs of knowledge-workers, from writing (Noy and Zhang 2023), creative ideation (Epstein et al. 2023) and acquiring knowledge (Jo and Park 2023), through to the study of science (Fauzi et al. 2023) and programming (Peng et al. 2023). Thus, about 80% of workers could have at least 10% of their work tasks affected, while around 20% could have at least 50% of their tasks affected (Peng et al. 2023). These individual productivity gains should benefit company performance, although there is little evidence to date to confirm this notion (e.g., Chu 2023; Budhwar et al. 2023). It is also still unclear when these effects will materialise, but 25% of CEOs already expect a 5% reduction in employee headcount in 2024 (PricewaterhouseCoopers 2024), due to efficiency gains engendered by generative AI. Even on a global level, AI will have a tremendous impact. Researchers propose its potential to contribute to the accomplishment of complex sustainable development goals in relation to society (no poverty, quality of education, etc.), the economy (decent work and economic growth, innovation and infrastructure, reduced

inequalities, etc.) and the environment (climate action, life below water, use of land) is exceptional (Vinuesa et al. 2020).

Notwithstanding these positive opportunities for individuals, organisations and society, this thesis also addresses the darker side of AI by pointing out and addressing worrying and potentially negative ethical impacts, which are especially relevant in terms of high-risk decision-making where people's careers, futures and lives are directly affected. Human Resource Management (HRM) is one of those high-risk decision-making areas confronted with the unresolved paradoxical duality of positive and negative effects on people's work autonomy and career opportunities as a result of AI and ML (Meijerink and Bondarouk 2021; Charlwood and Guenole 2022; Edwards et al. 2022).

In 2016, HR started using the term *Big Data* to examine in detail both the positive and dark sides of datafication at a granular level beyond traditional performance measurement (Angrave et al. 2016). Thereafter, the field gradually adapted to new technological advances by using the terms *HR analytics* (Marler and Boudreau 2017), *people analytics* (Tursunbayeva et al. 2018), *algorithmic HRM* (Meijerink et al. 2021), *HRM algorithms* (Meijerink and Bondarouk 2021), *ML in HRM* (Hickman et al. 2021) and *AI in HRM* (Chowdhury et al. 2022). Although these partially synonymous terms have different focus areas and emphases in their definitions, they essentially refer to similar methods, approaches and practices. In this thesis, a broad focus is placed on AI, which has been defined as "systems that behave like humans or perform human tasks", following the early work in this field (Turing 1950). In particular, this work focuses on ML, a subfield of AI, defined as "the ability of algorithms to learn without being explicitly programmed" (Samuel 1959). Since AI, ML or algorithms in general cannot be clearly separated from each other, the terms are used similarly or synonymously in many contexts. This thesis follows the definition provided by Meijerink et al. (2021), in that algorithmic HRM can encompass AI, ML and other statistical methods from analytics.[1]

> **!** *"Algorithmic HRM is the use of software algorithms that operate on the basis of digital data to augment HR-related decisions or to automate HRM activities". (Meijerink et al. 2021, p. 2547)*

As we shall discuss in the following sections, the duality of the bright and dark sides of AI in HR, and high-risk decisions in general, is receiving a lot of attention from politics, business and society. The demand for technologies that ensure the responsible use of AI is therefore a focal point of the current discussion on technological developments in a field in which – as in generative AI – high expectations are currently being placed (see *Figure 1*).

[1] Please note that the terms 'AI in HRM', 'algorithmic HRM' and 'ML-based HR Analytics' are used synonymously in the three papers of this thesis. Different terms are used in the three articles of the cumulative thesis to take into account the specific background and context of the publishing journals.

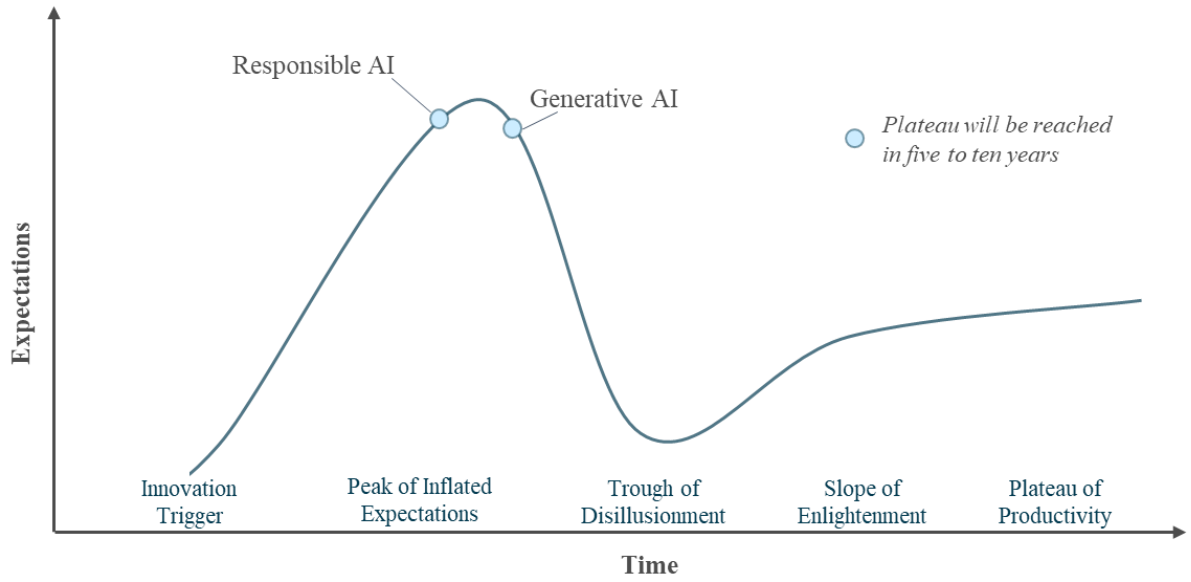


Figure 1: Selected technologies in the Gartner Hype Cycle ™ for AI, adapted from Gartner (2024).

The overarching goal of this study is to help shape the future of algorithmic HRM on the positive bright side of the continuum and realise the expected impact of responsible AI technologies. To achieve this aim, this cumulative thesis contributes to overcoming obstacles to realising the potential that enables the organisational value of ML algorithm adaptation while ensuring key requirements for responsible AI for algorithmic HRM, such as transparency and fairness. In doing so, it addresses the blind spots in HRM research identified by Lamers et al. (2024) by assuming algorithmic HRM as a given technological black box "out there in the world," without considering the immersion of human actors, embedded data input and the design or implementation processes of models using field-based methods.

1.1.1 The bright side of AI and algorithmic HRM

To date, evidence on the above-promised impact of (generative) AI on organisational performance is limited (Budhwar et al. 2023). However, looking at more established data-oriented approaches to HRM, management science finds that the adoption of analytics improves organisational performance (Aral et al. 2012). Recent research therefore concludes that HR analytics and algorithmic HRM capabilities are a key resource for a strategically relevant HRM discipline that contributes to higher organisational performance (Angrave et al. 2016). As a result, the traditional HR department, which was based more on qualitative approaches, has developed into one of the most analytical functions in the company in recent years (Davenport 2019). McCartney and Fu (2022) explain the impact of algorithmic HRM on organisational performance based on the theories of evidence-based management, dynamic capabilities and the resource-based view (see *Figure 2*).

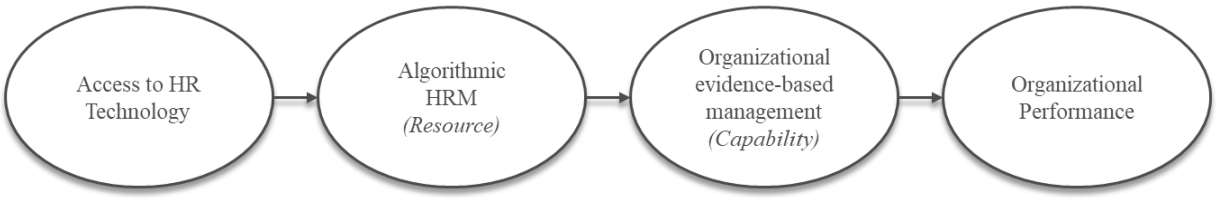


Figure 2: Chain model for the impact of algorithmic HRM on organisational performance, extracted from McCartney and Fu (2022)

From a technological perspective, algorithmic HRM options include language-related techniques (e.g., natural language processing and speech recognition), thought-related techniques (e.g., solution search and mathematical optimisation) and knowledge-related techniques (e.g., knowledge representation, retrieval and discovery) (Strohmeier and Piazza 2015). From a managerial perspective, these descriptive and diagnostic techniques enable numerous algorithmic HRM applications in all core HRM processes (selection, training, workforce planning, appraisal and compensation) (Margherita 2021;Meijerink and Bondarouk 2021).

Beyond the advantages of classical analytics, the application of ML in algorithmic HRM promises the *objective* creation (free from human bias) of *new* knowledge (not conceived or anticipated by humans) in an *efficient* way (exceeds the speed and abilities of humans) (van den Broek et al. 2021). The reason for this is that ML outperforms humans in predictive tasks, especially in highly complex scenarios, due to its ability to analyse large amounts of data efficiently (Agrawal et al. 2018). ML either automates repetitive and non-judgmental tasks or augments non-routine decision-making processes with predictions based on tasks that need to be performed by humans (Budhwar et al. 2022). Therefore, it has been emphasised that ML contributes to a number of HRM objectives influencing employee experiences, including (1) promoting diversity, (2) meaningfulness of work, (3) job autonomy, (4) job satisfaction and (5) motivation (Daugherty et al. 2018; Budhwar et al. 2022; Chowdhury et al. 2022). Similarly, on an organisational level, ML and AI contribute to several desired outcomes, including (1) acquiring, developing and retaining talent, (2) process efficiency, (3) customer satisfaction, (4) sustainability goals, (5) brand reputation and (6) HR cost efficiency (Chowdhury et al. 2022; Budhwar et al. 2022; Malik et al. 2023). In the most advanced and comprehensive applications, ML algorithms prescriptively take over most or all relevant HRM decision-making processes, leading to digital platforms that enable, among other things, short-term contracts or freelancing, i.e., so-called 'gig' ecosystems. Examples in this regard include successful intermediary platform firms that connect requesters (i.e. organisations or consumers) with on-demand gig workers in industries such as transportation (e.g., Uber), cleaning (e.g., Helpling), household do-it-yourself (e.g., TaskRabbit), food delivery (e.g., Just Eat) and programming (e.g., Clickworker) (Meijerink and Keegan 2019). Interestingly, all types of analytics can benefit from the technological advances of generative AI, as it accelerates and drives the adoption of a data-driven culture, which in turn is a key enabler for algorithmic HRM (Davenport and Bean 2024).

In summary, the introduction of algorithmic HRM enables evidence-based management and promises more accurate, objective and effective decision-making processes that can mutually benefit both a company and its employees.

1.1.2 The dark side of AI and algorithmic HRM

Contrary to these bright promises of AI, critical authors envision a dystopian future dominated by discriminatory and unfair algorithms – the dark side of AI – in which decision-makers who are unfamiliar with the algorithms' capabilities are deceived or misled into making wrong decisions, or could even use them as an excuse to make discriminatory decisions (Pasquale 2015; O'Neil 2016). For example, research has raised concerns about the widespread stereotypes that lead to gender bias when analysing textual data using Natural Language Programming (Bolukbasi et al. 2016). The literature points out that even if ML developers try to solve these problems, not taking the social context into account is a risk that cannot be ruled out due to the abstraction and simplification used by the algorithms building models of the real world (Selbst et al. 2019).

Algorithmic HRM can have a significant impact on individuals' lives and careers and has therefore raised concerns, particularly because ML models can incorporate non-traceable and unfair biases (Cheng and Hackett 2021). Unintended and far-reaching consequences of malfunctioning algorithms, such as unintentionally perpetuating and amplifying societal inequalities, are already well documented. Amazon, for instance, has stopped using and developing its ML-based recruitment system because it unintentionally perpetuates gender, ethnic or socio-economic inequalities (Meyer 2018; Dastin 2022). Similarly, Facebook's ad-targeting algorithms disadvantaged women in job postings because they were less likely to see ads from companies that hire predominantly male employees (Lambrecht and Tucker 2019; Teodorescu et al. 2021). Research also shows that controlling employees through an opaque algorithm (e.g., by replacing them or rewarding them on completion of tasks) negatively impacts the employee experience by causing insecurity, frustration and stress (Kellogg et al. 2020).

The above examples of the dark side of AI are based on a worrying prioritisation of economic over social values. Driven by technology hype, the algorithmic HRM industry focuses on commercial priorities, often neglecting the ethical and social impacts of its work (Charlwood and Guenole 2022). Both private and public private organisations prefer to focus on the benefits associated with positive business cases and initiate change processes in the early stages of AI implementation to increase adoption, but they give little to no consideration to ethical considerations such as accountability, fairness and transparency (Neumann et al. 2022). The dark side of AI and algorithmic HRM is exacerbated by the growing possibilities of generative AI tools, which “create continuing uncertainty for workers, expanding their business applications, while heightening risks related to well-being, bias, misinformation, context insensitivity, privacy issues, ethical dilemmas, and security” (Budhwar et al. 2023, p. 607).

In summary, the introduction of algorithmic HRM has the potential to lead to worrying and potentially negative ethical implications that may reinforce or spread injustice, discrimination and unfairness, which in turn would mutually harm both organisations and their employees.

1.2 Background and contextual framework

1.2.1 High-risk decision-making and foundations of responsible AI

In recent years, these negative consequences of the dark side of AI have gained increasingly more attention. Political debate and society are concerned about the impact of generative AI and ML algorithms in general use for modern weapons, as well as fake news or other consequences such as privacy concerns, technostress and addiction (Mikalef et al. 2022, p. 257). Being aware of possible consequences, the European Union began work on comprehensive AI legislation, namely the "European Union AI Act" (EU AI Act). Based on ethical guidelines and the suggestions of experts from science and practice (AI HLEG EU 2019), the legislation aims to follow a risk-based approach to classifying AI applications according to potential harm to individuals (European Parliament 2024). *Figure 3* provides examples of each of the four risk categories.

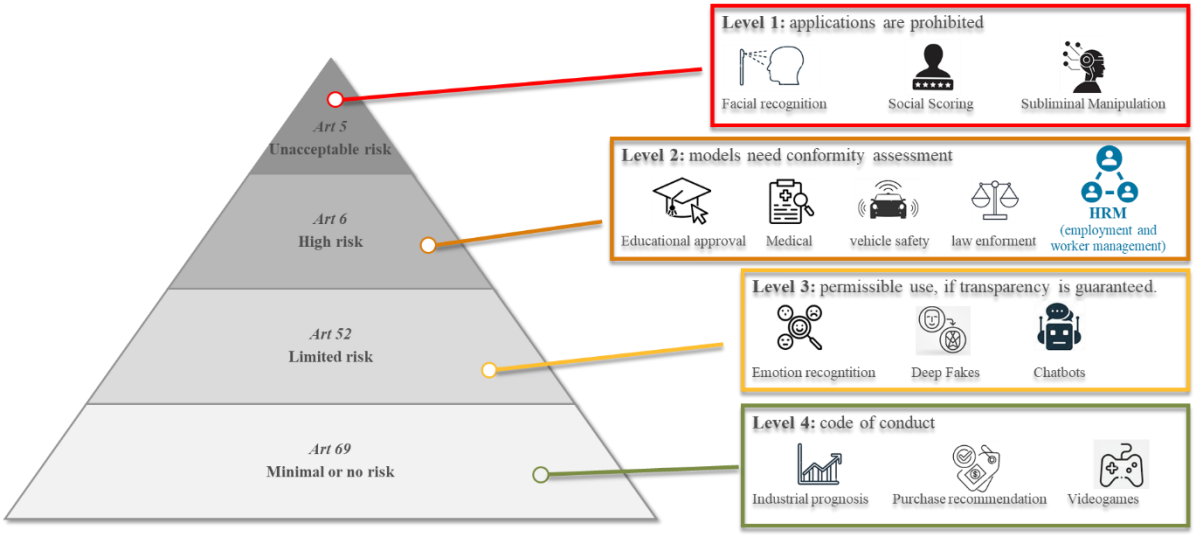


Figure 3: A risk-based regulation pyramid according to the EU AI Act, adapted from Díaz-Rodríguez et al. (2023)

Following the official European Union Artificial Intelligence Act (AI Act), adopted by the European Parliament on March 13, 2024, the HRM system relevant to this thesis is classified as a high-risk AI application in Annex III, under the category of 'employment, worker management, and access to self-employment.' (European Parliament 2024). Accordingly, for high-risk applications in Annex III, enhanced internal audits are required to assess conformity, albeit an external party for approval audits is not mandatory. For instance, organisations are asked to (1) document the intended purpose of the AI system in question, (2) provide detailed user instructions, (3) disclose the methods used to develop the system and (4) justify the critical design choices made by the provider (Mökander et al. 2022). In addition, the AI HLEG EU (2019) has identified seven requirements for high-risk applications that need

to be continuously assessed and addressed throughout the lifecycle of AI systems to ensure responsible use (see *Figure 4*).

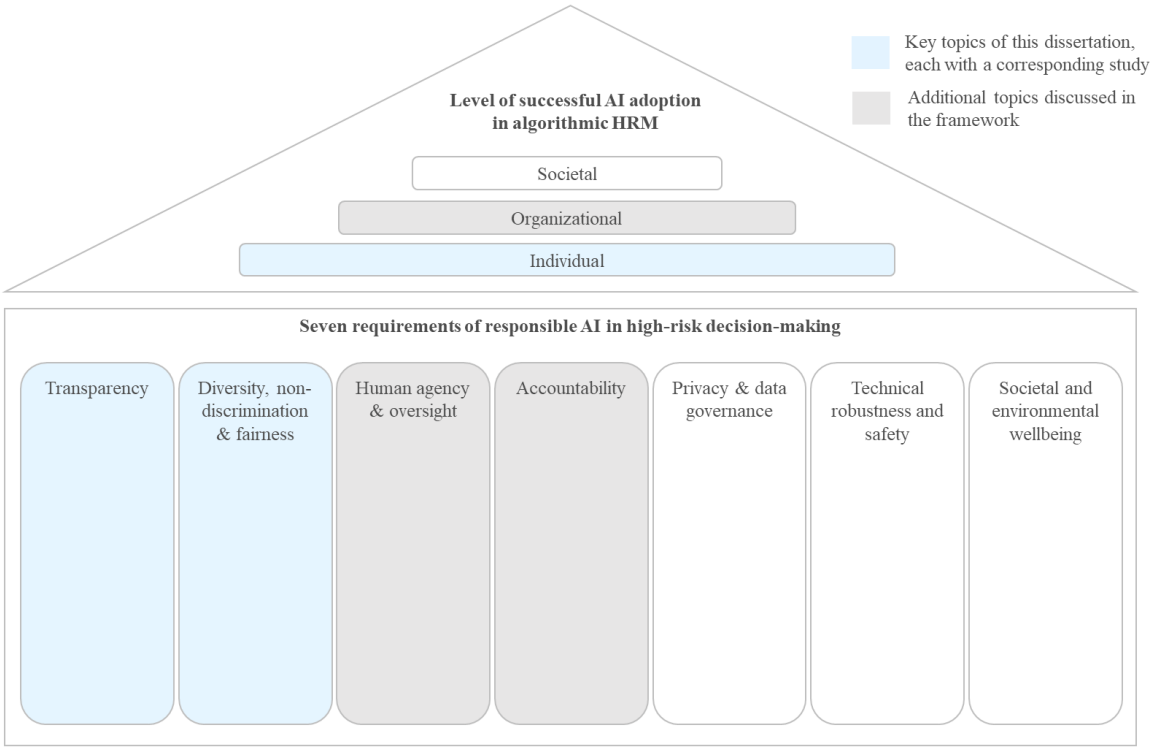


Figure 4: The responsible use of AI and the adoption of AI in HRM, and how this work contributes to this The overview includes the seven requirements when responsibly using AI in high-stakes decision-making, adapted from AI HLEG EU (2019).

As illustrated in *Figure 4*, this thesis does not address all seven requirements of responsible AI. Instead, two of the three embedded studies are dedicated to the most important requirements for responsible use, specifically in algorithmic HRM (Edwards 2022), namely transparency and diversity, non-discrimination and fairness. Furthermore, three additional requirements (human agency & oversight, accountability, privacy & data governance) are touched upon, especially in study 3, and briefly discussed later in the comprehensive discussion section.

It should be noted that 'transparency' is of paramount significance in this thesis. ML transparency is proposed by computer scientists in the XAI field as a fundamental precursor to responsible AI. A specific paper motivating this thesis originates from Arrieta et al. (2020), who provide a comprehensive overview of why, how and when ML transparency contributes to responsible AI. The proposition of the central importance of ML transparency is also based on extensive literature from managerial and HRM science that emphasises the need for ML transparency and explainability as an essential foundation for the responsible and value-added use of algorithmic HRM (Leicht-Deobald et al. 2019; Tambe et al. 2019; Gal et al. 2020; Glikson and Woolley 2020; Kellogg et al. 2020; Choudhury et al. 2021; Cheng and Hackett 2021; Langer and König 2023; Meijerink et al. 2021; Budhwar et al. 2022; Chowdhury et al. 2022; Bauer et al. 2023). While transparency is the focus requirement of Study 1, it is also central to the

ensuing two studies. The next section therefore provides a basic understanding of the technical background relevant to all three studies included herein.

1.2.2 The technical foundation of Explainable AI

In response to calls from both academia and industry for greater ML transparency, researchers in the field of information systems (IS) have begun to develop frameworks and technical explanatory approaches – also referred to as 'XAI methods' – to support the responsible, human-centred and trustworthy use of AI (e.g., Adadi and Berrada 2018; Arrieta et al. 2020; Molnar 2022; Ali et al. 2023).[2] The methods address the problem whereby ML algorithms learn by building their representation of a decision without considering human understanding, thereby escaping human understanding and interpretation (Burrell 2016; Kellogg et al. 2020). Knowledge about algorithms is insufficient to comprehend the inner-working of ML models, as the amount of information contained in built-in logics causes paradoxical effects such as information fatigue and meaninglessness (Gal et al. 2020). However, using XAI helps understand how an ML system decides, predicts and performs tasks, thus enabling a user to understand the limitations of its rationale (Shin 2021; Chowdhury et al. 2022; Zirar 2023). The fundamental aim of XAI is to create transparent models for predictions and recommendations in order to enable interpretability and make results explainable (see *Figure 5*).

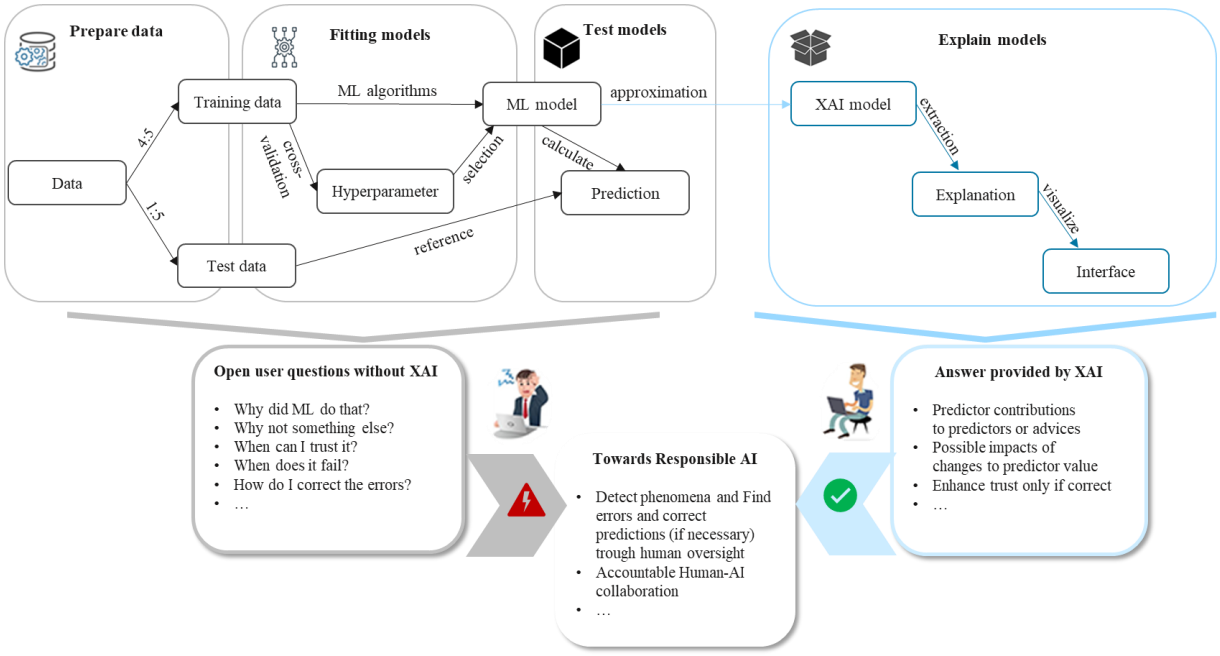


Figure 5: Fundamental concepts and value proposition of XAI compared to non-explained ML models in terms of user perception, adapted from Yuan et al. (2021) and Gunning and Aha (2019).

In contrast to traditional statistical algorithms that aim to develop inherently interpretable models to extract knowledge, such as linear regression, classification trees or Bayesian models (transparency-by-

[2] A complete overview of the numerous technical XAI solutions available is beyond the scope of this dissertation. For a more comprehensive overview, please refer to Adadi and Berrada 2018; Arrieta et al. 2020; Molnar 2022; Ali et al. 2023.

design), recent advances in the computer science literature set out to restore at least some of the interpretability of complex 'black box' algorithms, such as neural networks or random forests, by simplifying approximations (Ribeiro et al. 2016). The biggest advantage of the latter so-called 'post-hoc' explanations involves preserving the high predictive performance (e.g., accuracy, low error rate) offered by black box models and thereby mitigating the trade-off between predictive performance and transparency (e.g., Adadi and Berrada 2018; Arrieta et al. 2020; Molnar 2022). In the studies in this cumulative thesis, three different popular post-hoc explanatory methods are applied:

1. **Pertubated Feature Importance (PFI)** extracts global feature importance using the random permutation strategy, whereby feature values are randomly shuffled. This breaks the relationship between the predictor and the true prediction target. Comparing the resulting model's accuracy with the model before perturbation makes it possible to determine the factor by which the accuracy of the model increases with respect to this predictor. It should be noted that this method only calculates importance and is not able to quantify an effect via which the predictor increases or decreases the prediction (Molnar 2022).
2. **SHapley Additive exPlanations (SHAP)** is a post-hoc XAI method separately computing the effect of each predictor on a specific prediction (Lundberg and Lee 2017). To do this, SHAP divides the complex ML model formula into several mathematical vicinities (Chowdhury et al. 2022). What distinguishes SHAP's explanations from other local (specific for one prediction) XAI methods is their additive nature, which results from the former's game-theoretical approach. In analogy to game theory in economic behaviour (Morgenstern et al. 2007), SHAP decomposes the probability of a prediction into SHAP values, which in turn quantify the increasing or decreasing contribution made by individual predictors (the rational agents). Finally, the sum of all the SHAP values is the difference between the specific prediction and average predicted probability (Lundberg and Lee 2017).
3. **Accumulated Local Effects (ALE)**, a global (generalising information of many/all predictions of a model) post-hoc XAI method, extracts non-linear predictor effects on cohorts or subgroups of employees (Apley and Zhu 2020). ALE describes how a single predictor influences the prediction (strength, positive/negative contribution) on average, considering all employees in the respective cohort (Apley and Zhu 2020). Compared to alternative visualisation techniques on a global level (e.g., partial dependence plots), it is less vulnerable to unreliable extrapolation and is thus preferred when training ML on datasets with multicollinearities because it uses conditional distributions on each segment to calculate predictor effects (Apley and Zhu 2020; Molnar 2022).

1.3 Outline

1.3.1 Characterisation of the overarching case study

This cumulative thesis includes three studies, each one of which is based on an in-depth case study conducted in a German federal agency with more than 20,000 employees in the period between 2021 and 2024. This thorough public sector study approach provides a context in which high requirements for the individual adoption of algorithmic HRM can be investigated, and commercial secrecy is not a concern (Desouza et al. 2020, p. 206; Busuioc 2021, p. 826). The three studies include illustrative (examining the implementation and outcomes of innovative practices, particularly in studies 1 and 2) and exploratory (examining how and why the practices were adopted, particularly in study 3) aspects of the case study methodology (Smith 2019, pp. 204–205). In the following, the motivation behind and background of the organisational setting are provided.

Due to the older age structure of the workforce, the low number of qualified applicants and the anticipation that the workload will steadily increase over the next few years, the organisation is facing an urgent shortage of employees. The aim is to develop an ML model that predicts individual voluntary turnover (excluding age-related reasons and termination on the part of the employer) probability for each employee. Possible contributions and benefits of the ML-based employee turnover prediction project on the overall HRM goals may primarily be to (1) proactively identify the number of departing employees, in order to potentially initiate retention measures, (2) understand the reasons why certain subgroups (‘personas’) leave the organisation, (3) anticipate shortages in certain skills and (4) increase the precision of operational workforce planning to reduce staffing bottlenecks. While the organisation frequently uses descriptive analytics based on advanced dashboarding tools, as well as sporadic diagnostic regression-based analytics, this project is the first initiative to incorporate complex ML models and predictive analytics approaches. As employees can be directly affected by decisions derived from the model, this ML-based algorithmic HRM application falls under the high-risk categorisation according to the EU AI Act (see *Figure 3*).

The ML models developed herein are based on a uniquely collected dataset. This comprehensive dataset includes structured (tabular) information about 680,000 data points from a German federal agency. Each row represents an employee in combination with a specific month and contains a dichotomous variable about whether they left the company in the six months following the observed month. The dataset includes 27 predictors, most of which (19) originated from internal HR databases. These include work-related predictors such as commuting distance, sick days, salary, salary increases in recent years, seniority and others. Demographic data such as gender, age, number of children and education level are also included. In addition, data on global employee satisfaction and some data of external sources were integrated into the dataset. When collecting the anonymised data, a data protection procedure was used based on established data agreements and anonymisation procedures within the organisation.

The target variable of voluntary employee turnover is imbalanced, meaning that voluntary turnover is relatively rare (rate about 1%) in this dataset, which is not unusual in the public sector (e.g., Grissom et al. 2016, p. 243). However, it should be noted that the unbalanced class ratio makes prediction particularly difficult (Kuhn 2019). ML predictions are evaluated in an out-of-sample test dataset (1:5 split), and instead of treating the ML models as a black box, post-hoc explanations at local (employee-specific) and global (organisation-wide) levels are used to extract the predictors' effects. The confusion matrix is used to assess predictive accuracy and exemplary results of post-hoc explanations at the employee-specific level or as a communication tool and the basis for iterative development and troubleshooting for prediction errors. All steps (data preparation, model development, statistical tests and optimisation, tuning of hyperparameters, XAI methods, fairness evaluation, etc.) are implemented with the *R* programming language and an environment for statistical computing (R Core Team 2013), as well as the CARET (Classification and regression training) (Kuhn 2019) and IML (interpretable machine learning) (Molnar et al. 2018) packages. *Figure 6* characterises the most important properties of the ML models in red.

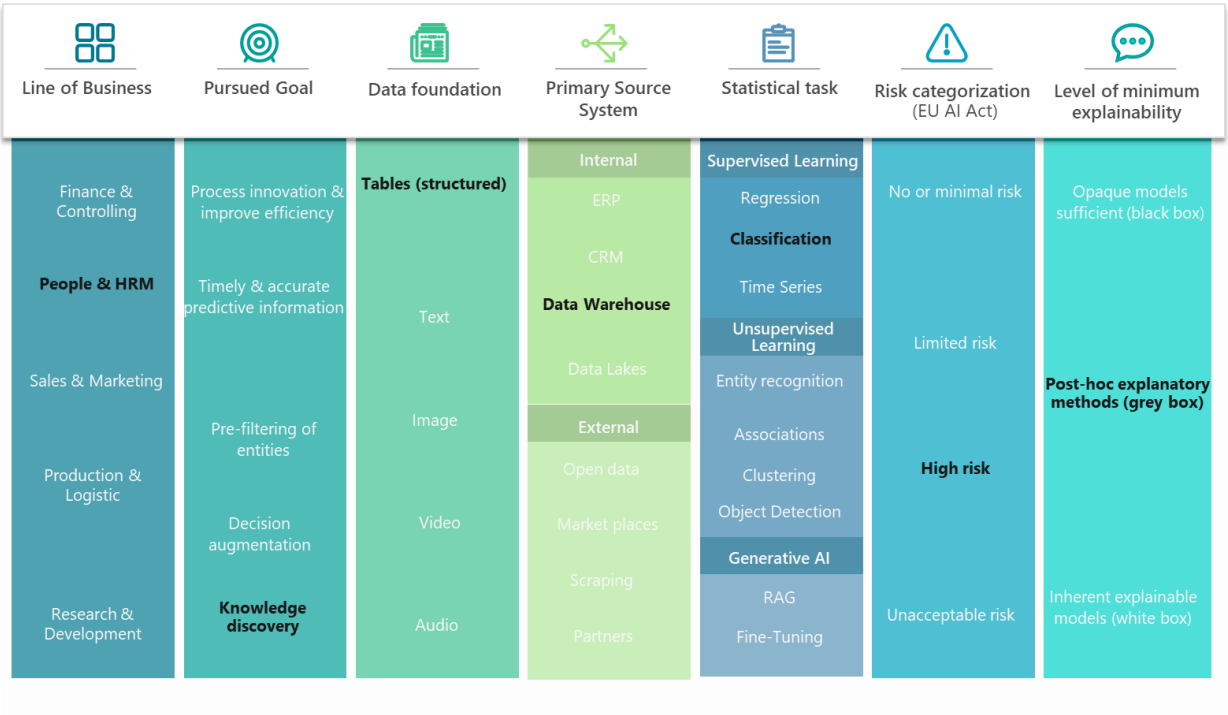


Figure 6: Characterisation of the ML model, which is the focus of the case study relevant in all three studies.

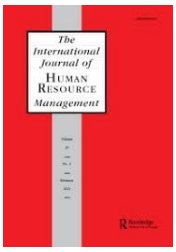
Besides the dataset and ML models, the case studies' information foundation includes a number of resources such as internal documents, e-mails, meeting notes, presentations following the implementation of a modern cloud-based analytics IS for workforce planning and forecasting that includes the above ML model. In addition, for study 3, interviews were collected from employees from several HR departments (HR Analytics, HR planning & forecast, HR management accounting & control, employee development) as well as managers of teams (10-20 employees) or departments (20-100

employees). Last, documented and openly available self-regulation is a valuable resource. Based on the EU AI Act, the "German Network Artificial Intelligence in Labour and Social Administration" (2022) provides in-depth clarification of the use of ML in algorithmic HRM in German federal authorities and sets out the same seven requirements before the law comes into force.[3] Overall, the case study provides an interesting and rich empirical field setting that can be used to provide realistic and ethnographic insights that are currently lacking in the literature on algorithmic HRM (van den Broek et al. 2021; Charlwood and Guenole 2022; Kelan 2023).

1.3.2 Overview of the three studies

The three studies contained in this cumulative thesis will apply XAI to high-risk decision-making in HRM from different perspectives. (see *Figure 4*).

STUDY 1	**"Machine learning with real-world HR data: Mitigating the trade-off between predictive performance and transparency"**
Status	published, DOI: 10.1080/09585192.2024.2335515 [4]
Co-Authors	Hülter, Svenja Marie; Tekieli, Michael
Keywords	Algorithmic HRM, Human Resource Analytics, Machine Learning Transparency, Explainable AI, Voluntary Employee Turnover Prediction
Journal	*The International Journal of Human Resource Management*
Metrics	VHB-Ranking: **A**[5], Impact Factor 4.9 (**Q1**) & Cite Score 11.7 (**Q1**)[6], H-Index 139 (**Q1**)[7]
Description	Study 1 provides evidence for the existence of a trade-off between predictive performance and transparency in ML-based algorithmic HRM. This trade-off has not been found in similar empirical studies, which is important because it motivates the application of post-hoc XAI methods. Only when transparency-by-design approaches using simpler algorithms fail may the approximation and simplification of post-hoc explanations be justified. The paper thus concludes with a nuanced view of the positive and negative aspects of post-hoc explanatory methods in high-risk HRM applications, answering important questions on the transparency requirement for responsible ML.



[3] Self-regulation is mainly based on the AI HLEG EU 2019 but also by the German Data Ethics Commission, the German Bundestag's AI Study Commission, the German government's AI strategy, the German Data Protection Conference and the OECD's Council on Artificial Intelligence (German Network Artificial Intelligence in Labour and Social Administration2022).

[4] https://www.tandfonline.com/doi/full/10.1080/09585192.2024.2335515

[5] VHB Rating 2024 https://vhbonline.org/fileadmin/user_upload/VHB_Rating_2024_Area_rating_PERS.pdf

VHB Rating 2015: B https://vhbonline.org/vhb4you/vhb-jourqual/vhb-jourqual-3/gesamtliste

[6] https://www.tandfonline.com/action/journalInformation?show=journalMetrics&journalCode=rijh20

[7] https://www.scimagojr.com/journalrank.php?category=1407

STUDY 2	**"Towards fair Human Resource Analytics: Introducing a sociotechnical framework"**
Status	accepted
Co-Authors	*None*
Keywords	Human Resource Analytics, Algorithmic Fairness, Explainable AI, Sociotechnical Theory
Conference	*3rd EIASM Workshop on People Analytics & Algorithmic Management (PAAM)*[8]
Description	Study 2 moves on to a second central requirement of responsible AI: diversity, non-discrimination & fairness. The study aims to bridge the gap between social science and information science and contributes to understanding how sociotechnical systems designed in ML fairness assessments help mitigate the amplification of adverse impacts against minorities through algorithmic HRM. The main contribution is the introduction of a holistic framework specifying processes, responsibilities and interconnections between automated bias mitigation approaches and human judgement.

STUDY 3	**"Exploring the individual adoption of Human Resource Analytics: Behavioural beliefs and the role of Machine Learning characteristics"**
Status	published, DOI: 10.1016/j.techfore.2024.123709)[9]
Co-Authors	Hülter, Svenja Marie; Ertel, Christian
Keywords	*Human Resource Analytics, Machine Learning Adoption, Explainable AI, Theory of Planned Behaviour, Employee Turnover Prediction*
Journal	*Technological Forecasting and Social Change*
Metrics	VHB-Ranking: **B**[10], Impact Factor 12.9 (**Q1**) & Cite Score 21.3 (**Q1**)[11], H-Index 179 (**Q1**)[12]
Description	Study 3 addresses an important obstacle in trying to benefit from responsible AI, namely the adoption of ML at an individual level. The reason for this is documented in interdisciplinary research and indicates a lack of trust and acceptance of ML recommendations or predictions. Based on a focused interview method and the theory of planned behaviour, the study contributes to the literature by identifying relevant beliefs and experiences that influence the intention to adopt algorithmic HRM. Furthermore, the study provides insights into the impact of responsible AI on adoption by linking ML characteristics such as transparency, automation and fairness to the intention to adopt employee turnover predictions.

Table 1: Overview of the three studies contained in this cumulative thesis. Metrics according to 09.01.2025.

[8] https://www.eiasm.org/frontoffice/event_announcement.asp?event_id=1719%20

[9] https://www.sciencedirect.com/science/article/pii/S0040162524005079

[10] VHB Rating 2024 https://vhbonline.org/fileadmin/user_upload/VHB_Rating_2024_Area_rating_TIE.pdf

VHB Rating 2015: B https://vhbonline.org/vhb4you/vhb-jourqual/vhb-jourqual-3/gesamtliste

[11] https://www.sciencedirect.com/journal/technological-forecasting-and-social-change/about/insights

[12] https://www.scimagojr.com/journalsearch.php?q=14704&tip=sid

2 Main section (cumulative)

2.1 Study 1 | Machine learning with real-world HR data: Mitigating the trade-off between predictive performance and transparency

Abstract

ML algorithms offer a powerful tool for capturing multifaceted relationships through inductive research to gain insights and support decision-making in practice. This study contributes to understanding the dilemma whereby the more complex ML becomes, the more its value proposition can be compromised by its opacity. Using a longitudinal dataset on voluntary employee turnover from a German federal agency, we provide evidence for the underlying trade-off between predictive performance and transparency for ML, which has not been found in similar human resource management (HRM) studies using artificially simulated datasets. We then propose measures to mitigate this trade-off by demonstrating the use of post-hoc explanatory methods to extract local (employee-specific) and global (organisation-wide) predictor effects. After that, we discuss their limitations, thereby contributing to the literature on ML in HRM by providing a nuanced perspective on the circumstances under which the use of post-hoc explanatory methods is justified. Namely, when a "transparency-by-design" approach with traditional linear regression is not sufficient to solve HRM prediction tasks, the translation of complex ML models into human-understandable visualisations is required. As theoretical implications and for the advancement of HRM research, this paper suggests that we can only extensively understand the multifaceted HR phenomena that provide us with real-world data if we incorporate ML-based inductive methods. However, we argue for a joint application with traditional deductive methods, as the interpretation of potentially over-approximating complex ML models with post-hoc explanatory methods may sometimes lead to inaccurate or misleadingly simplified results.

Introduction

ML algorithms are increasingly used in HRM research to reveal and explain multifaceted phenomena beyond linearity from data, known as an "ML-based inductive" research method (for an example relating to employee turnover prediction, see King 2016; Rombaut and Guerry 2018, 2021, Choudhury et al. 2021, Erel et al. 2021, Speer 2021, Yuan et al. 2021, Chowdhury et al. 2022). This method promises to objectively translate large amounts of raw data into new information provided by otherwise overlooked, complicated interactions between predictors with a level of efficiency not achieved by humans (van den Broek et al. 2021). Generally, this explanatory approach requires neither prior assumptions nor explicit hypotheses, which opens up various opportunities for HR research and practice (Choudhury et al. 2021). Thus, the ML-based inductive research method enables the investigation of multifaceted relationships, to gain exploratory insights and develop theory (Putka et al. 2018; Cheng and Hackett 2021).

However, a challenge arises due to the complexity of ML algorithms, i.e. the disadvantage of not providing an easily understandable mathematical formula (Kellogg et al. 2020; Erel et al. 2021). High ML model complexity occurs when an ML algorithm capable of modelling flexible functions beyond linearity (neural networks, random forest, etc.) is applied to a large dataset with multifaceted relationships (Choudhury et al. 2021). Complex models achieve high predictive performance, which is the extent of the ML model to solve a prediction task, depending on its context and goal, with different statistical measures (e.g. root-mean-squared error for regression, accuracy for classification). Nevertheless, complex models remain opaque, in that (a) predictors are not understandable (e.g. coming from a complex system itself), (b) relationships between predictors and predictions are hidden and (c) no explanation for a specific prediction is given (Burrell 2016; Langer and König 2023).

Simply, there are two sides to a continuum: opacity is a lack of understanding regarding the inner workings of the ML model, while transparency is understanding the relationships between predictors and predictions (Langer and König 2023). To ascertain the position on this continuum between transparency and opacity, ML algorithm selection is a pivotal determinant. For example, a linear regression model can be considered as inherently transparent, also known as the "transparency-by-design" approach, because the mechanisms of the mathematical formulas and their parameters, which affect the relationships between predictors and predictions, are interpretable. Generally, it is known that the trade-off between predictive performance and transparency in ML models goes hand in hand with their complexity (Arrieta et al. 2020, p. 100), but it has not been empirically demonstrated or investigated in real-world HRM applications. Thus, we ask:

> **RQ1:** To what extent does ML in real-world HRM applications face trade-offs between predictive performance and transparency?

We select employee turnover prediction as a representative ML application in HR. This has also recently been investigated in two closely related studies (Choudhury et al. 2021; Chowdhury et al. 2022). We extend these studies by providing an empirical example of the trade-off between predictive performance and transparency in real-world data that was not found using artificially simulated datasets used in either study. Also, like these studies, we apply multiple post-hoc explanatory methods to mitigate the aforementioned trade-off. These methods aim to explain elements of a complex ML model while, such as the influence importance of predictors, maintaining its high predictive performance, which increases its transparency. However, additionally we critique the limitations and ask about the appropriate circumstances for using post-hoc explanatory methods:

> **RQ2:** Under what circumstances, may post-hoc explanatory methods be applied to understand the rationale behind complex ML model predictions?

By answering these research questions, we contribute to the literature by (1) empirically demonstrating and discussing the consequences of the trade-off between predictive performance and transparency, (2) providing a nuanced perspective when the use of post-hoc explanatory methods is justified due to lack of alternatives (3) and outlining the possibilities of revealing the multifaceted effects of complex ML model predictors combined with post-hoc explanatory methods. Thus, we respond to research calls to investigate when and how organisations can switch from opaque to transparent ML (Chowdhury et al. 2022, p. 25). As a theoretical implication, our study exemplifies the need for an inductive method based on ML, given the complexity of real-world HR phenomena that cannot be investigated to the same depth using traditional methods such as linear regression. However, caution is needed when interpreting and deriving implications, as the available post-hoc explanatory methods required for complex models can be misleading. Confirmatory studies with deductive evidence may therefore be complementarily necessary.

The paper is organised as follows. The second section reviews the literature on the multifaceted causes of employee turnover, frameworks for ML-based inductive research methods and the trade-off in ML complexity. The third section summarises the methodology used and presents the empirical dataset. The fourth section presents the results and applies three post-hoc explanatory methods to mitigate a trade-off. Finally, the fifth and sixth sections discuss post-hoc explanatory methods before concluding.

Literature Review

Turnover causes are multifaceted

Employee turnover prediction is selected as a representative ML application in HRM because turnover is generally an intricate process (Yuan et al. 2021) resulting from a series of possible sequences of events or ‘pathways’ (Russell and Sell 2012, p. 126). Consequently, the literature lists numerous predictors directly influencing employee turnover (e.g. Holtom et al. 2008; Rubenstein et al. 2018). This suggests

that turnover is more complex than the largely linear relationships previously studied. However, turnover causes are likely to be nonlinear, with the nature of these relationships between variables changing at different points (e.g. U-shaped), or heterogeneous with different relationships for different subgroups of employees. For example, Gray and Phillips (1994) find a U-shaped relationship between age and turnover, implying that turnover is high among young employees and decreases with age; thereafter, turnover gradually increases again, due to retirement. A negative correlation between tenure and employee turnover is widespread (Rubenstein et al. 2018). Furthermore, Grissom et al. (2016) find a negative relation between salary predictors (total/increase) and turnover in public administration, albeit up to a certain level, whereas turnover at the managerial level might work differently. Lin et al. (2021) emphasise that wage increases significantly reduce voluntary turnover. In addition, they examine the moderating effect on the relationship between wage increases and turnover, finding a significant negative relationship in this regard, albeit only for workers with longer tenure.

In summary, the sophistication of relationships leading to employee turnover is not only caused by linear relationships between several predictors and turnover, but also nonlinearity and heterogeneity in employee subgroups, which lead to mathematical interactions among predictors.

Increasing algorithm complexity

Multifaceted employee turnover prediction phenomena contradict implicit assumptions of linearity in traditional ordinary least squares models, thereby strengthening the rationale for using ML for prediction and knowledge (Erel et al. 2021). Recent literature presents frameworks and proposals for inductive research methods that introduce the principles of modern predictive models (Yuan et al. 2021) and common algorithms (Putka et al. 2018), to embedded ML in scientific methodologies. For example, ML algorithms used for employee turnover prediction include random forest (Choudhury et al. 2021; Speer 2021), extreme gradient boosting (Erel et al. 2021) and gradient boosting machines (King 2016). These so-called “ensemble” algorithms combine hundreds of heterogeneous trees, each of which automatically selects relevant predictors and their weights, groups data into relevant subregions and finds local dependencies (Putka et al. 2018). Similarly, increasing the number of neurons per layer or the number of layers in a neural network leads to what are known as deep learning algorithms that can achieve the same result (Vale et al. 2022).

To avoid system-based ML opacity resulting from increasing complexity, knowledge about algorithms is insufficient, as the amount of information causes information fatigue and meaninglessness (Gal et al. 2020). Complex ML algorithms learn by building their representation of a decision without considering human comprehension, thereby escaping human understanding (Burrell 2016, p. 10; Kellogg et al. 2020, p. 372). In contrast to traditional statistical and econometric approaches, non-parametric ML algorithms do not provide a representative mathematical formula that can be easily understood (Erel et al. 2021, p. 3229). However, some do provide other representations that are understandable to humans, such as

the rule system of classification trees or the conditional probabilities of naive bayes classifiers (Arrieta et al. 2020). This means, the decision between a more transparent, simple ML algorithm and a more complex, opaque one is more of a continuum due to the variety of algorithm options available (Langer and König 2023). Regardless of the specific algorithm, it is widely assumed that higher ML model complexity directly correlates with higher prediction performance, or expressed differently, solves the prediction task more effectively. Interestingly, this is not necessarily the case, as it depends on the specific prediction task and the flexibility required to approximate the underlying data, in particular the amount available and its distribution among the variables' values (Rudin 2019).

Challenges arising from algorithm opacity

HRM is a high-stakes decision-making environment because individuals are directly affected, and data-driven decisions have various ethical and legal consequences (Gal et al. 2020). To verify fairness and non-discrimination, input data, data analysis procedures and the links between results and conclusions must be transparent so that predictions can be validated. One reason is that correlations can be established but causal relationships are not implied. Opaque algorithms jeopardise the ability to test for adverse impacts by challenging differences between groups with existing evidence (Meijerink et al. 2021, p. 2549; Charlwood and Guenole 2022, p. 7), thus hampering unavoidable critical thinking about predictive composites, causal inferences and subgroup differences (Putka et al. 2018, p. 711).

ML transparency is important not only for practical applications, but also for researchers who want to understand, for example, what nonlinear effects independent variables have. Somers et al. challenge the linear assumptions of HR theory regarding employee well-being and demonstrate the use of complex ML models such as neural networks to detect nonlinearities. The tools used, such as three-dimensional visualisation, help uncover the existence of nonlinear phenomena but do not provide sufficient transparency to extract explicit relationships or to include more than two predictors (Somers et al. 2021). Yuan et al. (2021) use ML to identify the key predictors of employee turnover in a sample of SMEs. However, because of the algorithm's opacity, it is vague how the predictors affect turnover. For example, perceived unfairness is the most important predictor, but it is unclear at what point perceived unfairness triggers turnover.

In summary, it is essential to the success of ML in HR applications to (1) disclose how an algorithm makes a decision, (2) ensure the right to challenge an outcome and (3) provide expertise to address these challenges (Cheng and Hackett 2021). Ultimately, ML transparency is needed to increase trust in ML and question its responsible use (Chowdhury et al. 2022). In the absence of sufficient empirical examples, algorithm opacity is an important blind spot in algorithmic HRM research (Meijerink et al. 2021, p. 2550; Edwards et al. 2022, p. 5).

Tackling algorithm opacity with technical solutions

Research outside of HR has introduced methods to increase ML model transparency while maintaining high predictive performance, thus mitigating the aforementioned trade-off. XAI, also known as "Interpretable ML," is a rapidly evolving interdisciplinary research area offering multiple technical solutions (Arrieta et al. 2020; Molnar 2022). Moreover, XAI methods can be divided into either algorithmic models understandable to humans with appropriate knowledge (transparency-by-design), or methods using approximation to explain elements of complex – and thus opaque – models through simplified representations (post-hoc explanatory methods) (Arrieta et al. 2020; Langer and König 2023). Interestingly, documented applications of these technical solutions remain sparse in HR (Langer and König 2023), albeit the first examples are promising. Choudhury et al. (2021) demonstrate the use of global (enterprise-wide) post-hoc explanatory methods by applying feature importance methods and partial dependence plots to study nonlinear interactions between employee turnover and its predictors. Yakusheva et al. (2022) also use partial dependence plots to reveal an unexpected U-shaped relationship between higher staffing levels and the number of readmissions. Erel et al. (2021) provide an example of how to gain insights into opaque models of complex algorithms by quantifying the contribution of each predictor for director performance and turnover. Additionally, they demonstrate the existence of nonlinear and heterogenous relationships by breaking down the effects of predictors, using local post-hoc explanatory methods. In the HR literature, Chowdhury et al. (2022) use a local (employee-specific) post-hoc explanatory method for employee turnover prediction, called "local agnostic model explanations" (LIME).

Knowledge gap

These examples show that post-hoc explanatory methods can be applied to complex ML models, improving our understanding of employee turnover causes. However, the proposed frameworks do not specify under what circumstances these methods should be used. One reason for this is the lack of a connection with the trade-off between predictive performance and transparency, which is not found in studies using artificially simulated datasets (Choudhury et al. 2021; Chowdhury et al. 2022). Instead, Choudhury et al. note that complex ML models (e.g., neural networks, random forests) offer only 'small performance increases over the baseline logistic regression' and explain this marginal performance gain as 'meaningful interactions and nonlinearities among variables are only relevant for a small subset of the data' (Choudhury et al. 2021, p. 48). Likewise, but independently, Chowdhury et al. come to a similar conclusion, as they find 'no significant differences' in predictive performance between transparent and more complex algorithms (Chowdhury et al. 2022, p. 13). This warrants further investigation of the trade-off between predictive performance and transparency in practical, real-world HR datasets. The findings could serve as a basis for understanding the justified use of more complex, opaque ML algorithms and subsequent post-hoc explanatory methods.

Methodology

In turnover prediction, inductive research methods help analyse actual turnover in longitudinal data available through HR information systems (HRIS) (King 2016; Rombaut and Guerry 2018, 2021), also known as 'attrition modeling' (Speer 2021). Building on these frameworks, we introduce a complete process for ML model development with addressing ML opacity depending on prediction task complexity. The post-hoc explanatory methods approximate a complex ML model and extract various explanations that provide the transparency needed to question the logic behind predictions. *Figure 7* presents a schematic overview of an inductive research framework.

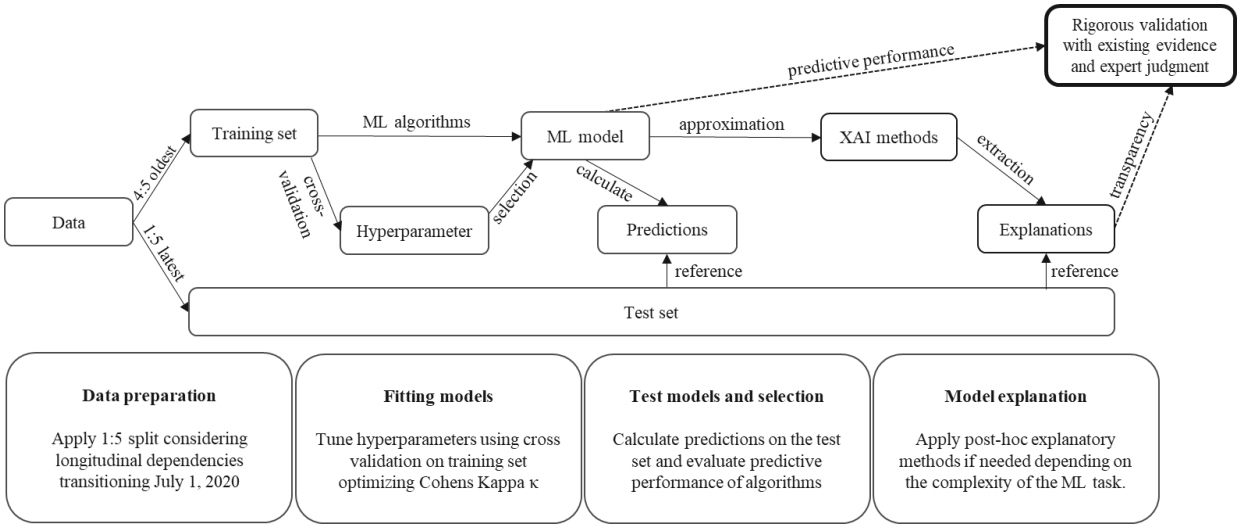


Figure 7: Inductive research process using machine learning with an out-of-sample test

Data preparation

We adapt Rombaut and Guerry's inductive approach to predict actual voluntary employee turnover, using data from HRIS, and extend it by using complex ML algorithms and external predictors (Rombaut and Guerry 2018). The direct effect (black arrow) is examined by bridging several implicit established constructs, such as organisational commitment and payment satisfaction, thus taking a complementary approach to survey-based methods (see *Figure 8*).

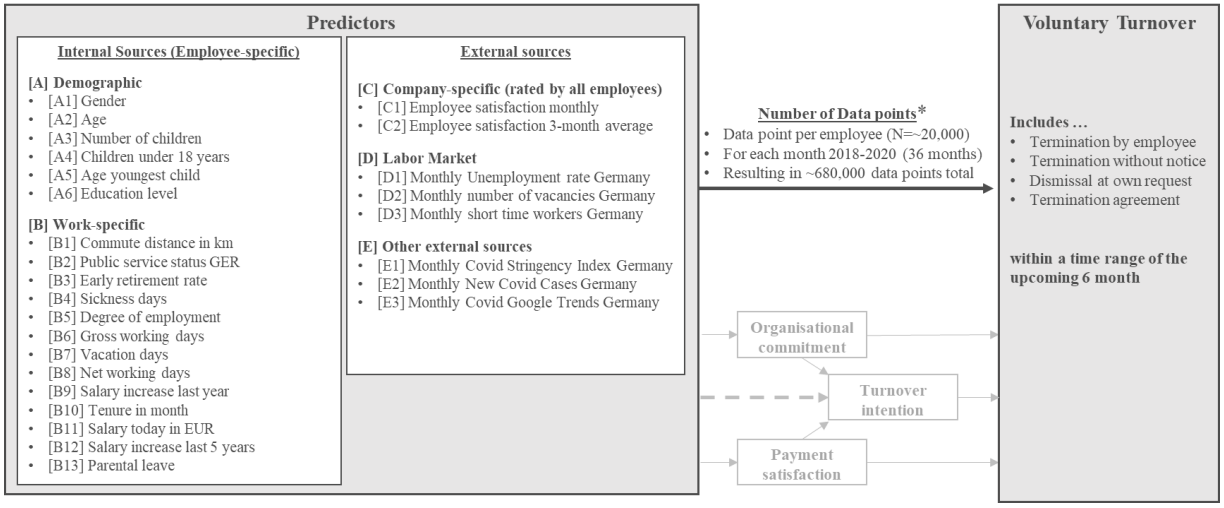


Figure 8: Acquired longitudinal data for historical voluntary turnover. The direct impact (black) is investigated in this study

The inductive ML methodology is based on a uniquely collected dataset of 680,000 data points from a German federal agency. Each row represents an employee in combination with a specific month and contains a dichotomous variable about whether they left the company in the six months following the observed month. This time horizon was chosen based on practical response time requirements for countermeasures; it is also consistent with other research (Speer 2021, p. 8). We used a data protection-approved process that builds on existing data agreements and anonymisation procedures during data collection.

The final dataset had 27 predictors (see *Figure 8*), most of which (19) originated from internal HR databases and are available for all 20,000 employees over 36 months. In addition, data on global employee satisfaction and external data were integrated into the dataset. *Table 2* provides a basic descriptive statistical overview of the relevant variables. Please note that the low voluntary turnover rate (below 1%) is not unusual in the public sector (e.g., Grissom et al. 2016, p. 243), thereby making prediction particularly difficult (Kuhn 2019).

Nominal Variables	N (679463)	Continuous Variables	Mean	Std. Dev.
A1_Gender		A2_Age	49.09	9.13
... woman	71,0%	A3_Number_of_children	1.13	0.94
... men	25.7%	A5_Age_youngest_children	20.46	10.09
…unknown	3.3%	B1_Commute_distance_in_km	15.28	17.62
		B3_Early_retirement_rate	6.08	17.12
A4_Children_under_18_years		B4_Sickness_days	2.07	4.73
... 0	72.6%	B5_Degree_of_employment	90.75	14.00
... 1	27.4%	B6_Gross_working_days	20.06	2.40
		B7_Vacation_days	2.33	3.55
A6_Education_level		B8_Net_working_days	15.34	5.78
... lower degree university	49.5%	B9_Salary_increase_last_year	0.16	0.22
... vocational training	46.7%	B10_Tenure_in_month	272.17	127.18
... higher graduation university	3.8%	B11_Salary_today_EUR	4068.64	889.97
		B12_Salary_increase_last_5_years	0.27	0.26
B2_Public_service_status_GER		B13_Parental_leave	0.06	1.11
... No	78,0%	C1_Overall_employee_satisfaction	3.36	0.63
... Yes	22,0%	C2_Employee_satisfaction_moving_average	3.50	0.32
		D1_Monthly_unemployment_rate_Germany	5.36	0.48
Voluntary turnover		D2_Monthly_number_of_vacancies_Germany	721.10	94.60
... No_Turnover	**99.2%**	D3_Monthly_short_time_workers_Germany	1134.91	1659.11
... Turnover	**0.8%**	E1_Monthly_Covid_strigency_index_Germany	19.67	29.15
		E2_Monthly_New_Covid_Cases_Germany	69835	181549
		E3_Monthly_Covid_Google_trends_Germany	74.70	116.21

Table 2: Descriptive statistical overview of available variables in the acquired dataset

Algorithm selection

Table 3 summarises the selected algorithms chosen from common options used for employee turnover prediction (Chowdhury et al. 2022) or other inductive research (Putka et al. 2018). The advantage (transparency vs. performance) is drawn from the computer science literature, which order algorithms according to their comprehensible representational capabilities (Arrieta et al. 2020, p. 90).

Algorithm	*R* base function/package	Advantage
Generalised Linear Model	*glm*	Transparency
Elastic Net Regression	*glmnet*	Transparency
Classification Tree	*rpart2*	Transparency
Naïve Bayes	*naivebayes*	Transparency
Random Forest	*ranger*	Performance
Extreme Gradient Boosting	*xgbDart*	Performance
Generalised Boosted Machine	*gbm*	Performance
Feed Forward Neural Network	*nnet*	Performance

Table 3: Selected ML algorithms with their primary advantage according to the transparency vs. performance trade-off

Fitting models on training data

We use three-way partitioning by initially training several algorithms and their hyperparameters to optimise Cohen's Kappa κ on training and validation data (cross-validation on first 24 months between January 2018 and June 2020). Cohen's Kappa κ is a performance indicator that expresses the chance-adjusted proportion of correctly predicted outcomes (Cohen 1960). It varies between zero and one, providing an interpretation similar to the traditional R-squared regression (Yakusheva et al. 2022, p. 315). By optimising Cohen's Kappa κ instead of other common evaluation methods (e.g., Accuracy, Receiver Operating Characteristic = ROC), we are able to achieve higher predictive performance across all algorithms when tuning hyperparameters, as it is better suited to address the challenge of an imbalanced dataset (Kuhn 2019).

Evaluate model performance on test data

Evaluating the predictive performance of an ML model is critical to ensure the model's suitability for providing valuable insights. We test the predictive performance of each algorithm with out-of-sample test data (last six months between July 2020 and December 2020).

Results

The predictive performance measure results are reported in Table 4.

Algorithm	Advantage	κ	*ROC*	*Precision*	*Recall*
Random Forest*	Performance	**0.26***	0.87	0.18	0.54
Extreme Gradient Boosting	Performance	**0.24**	0.85	0.18	0.36
Generalised Boosted Machine	Performance	**0.22**	0.87	0.18	0.34
Classification Tree	Transparency	**0.18**	0.68	0.23	0.15
Feed Forwards Neural Network	Performance	**0.16**	0.68	0.17	0.16
Generalised Linear Model	Transparency	**0.12**	0.77	0.23	0.09
Elastic Net Regression	Transparency	**0.12**	0.76	0.20	0.09
Naïve Bayes	Transparency	**0.07**	0.59	0.05	0.23

Table 4: Predictive Performance measures of all used algorithms on test data

We further compare the ML models with the highest predictive performance overall with the most successful algorithm (random forest), with the advantage of transparency (classification tree). The confusion matrices in Table 5 show the amount of cases on test data divided into positive and negative cases as well as correct and incorrect predictions.

Random Forest

		Reference	
		No Turnover	*Turnover*
Prediction	*No Turnover*	111,089 (96.60%)	513 (0.45%)
	Turnover	2,800 (2.43%)	**598** (0.52%)

Classification Tree

		Reference	
		No Turnover	*Turnover*
Prediction	*No Turnover*	113,324 (98.53%)	940 (0.82%)
	Turnover	565 (0.49%)	**171** (0.15%)

Table 5: Confusion matrices on test data of random forest (highest predictive performance overall) and classification tree (most successful alternative algorithm with advantage transparency)

The random forest model successfully identifies 598 (=52%) employee turnover cases. Overall, it provides a *fair* improvement over random guesses (κ = 0.26), according to common κ interpretation (Landis and Koch 1977). Measured by the recall (the ratio of true-positive cases to all positive cases) of 0.54, the model solves over half of the prediction task while providing sufficient precision (the ratio of true-positive cases to all positive predictions) of 0.18. Thus, given the challenge of highly imbalanced classes, due to the rarity of employee turnover cases, the difficult task of prediction is solved adequately, with an acceptable number of false-positive predictions.

The classification tree model successfully identifies 171 (=15%) cases and provides a *slight* improvement over random guesses (κ = 0.18). The model successfully identifies 427 fewer cases of employee turnover than the random forest model. While the precision of 0.23 is slightly higher, the recall of 0.15 is significantly lower than the random forest. Hence, the majority of turnover cases are not detected, suggesting that the classification tree cannot predict diverse causes of employee turnover (Russell and Sell 2012, p. 126). The same applies to the other transparent ML models with even lower overall performance ($\kappa < 0{,}12$). The low predictive performance of the two linear models (Generalised Linear Model, Elastic Net Regression) indicates that linear assumptions might only be valid up to a certain extent.

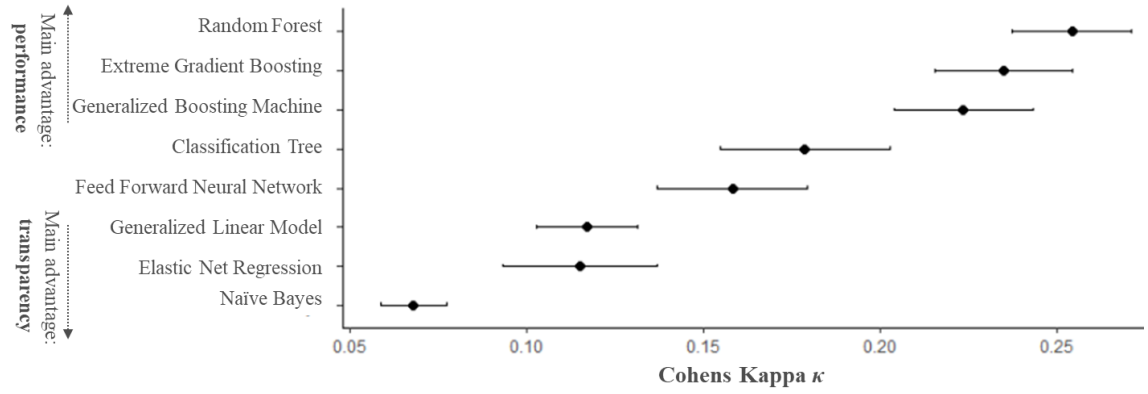


Figure 9: Predictive performance vs. transparency trade-off on test data (confidence interval 0.95)

Two findings emerge from these results. First, while all ML models can detect instances of employee turnover better than chance (κ > 0), they differ significantly in their predictive performance. Second, a trade-off between predictive performance and transparency is revealed, with more transparent

algorithms achieving lower predictive performance (see *Figure 9*). The only exception – classification trees achieve higher performance than feed forward neural networks – is due to tree-based methods tending to outperform neural networks for tabular data (Shwartz-Ziv and Armon 2022). Consequently, XAI's transparency-by-design approach is not sufficient for identifying various nonlinear or heterogenous causes for employee turnover as the ML model cannot predict most turnover cases.

Applying post-hoc explanatory methods to complex ML models

Thus, we apply three popular post-hoc explanatory methods as the remaining options to gain insights from the random forest model. The choice of method from among numerous alternatives is beyond the scope of this study, so we refer the reader to the technical literature (e.g. Molnar 2022). All three methods are implemented with the R package "IML" (Interpretable ML) (Molnar et al. 2018).

Global Feature Importance

Global feature importance is extracted using the random permutation strategy, which determines the factor by which the model's classification error increases when feature values are randomly shuffled, thereby breaking the relationship between the feature and the true outcome (Molnar 2022).

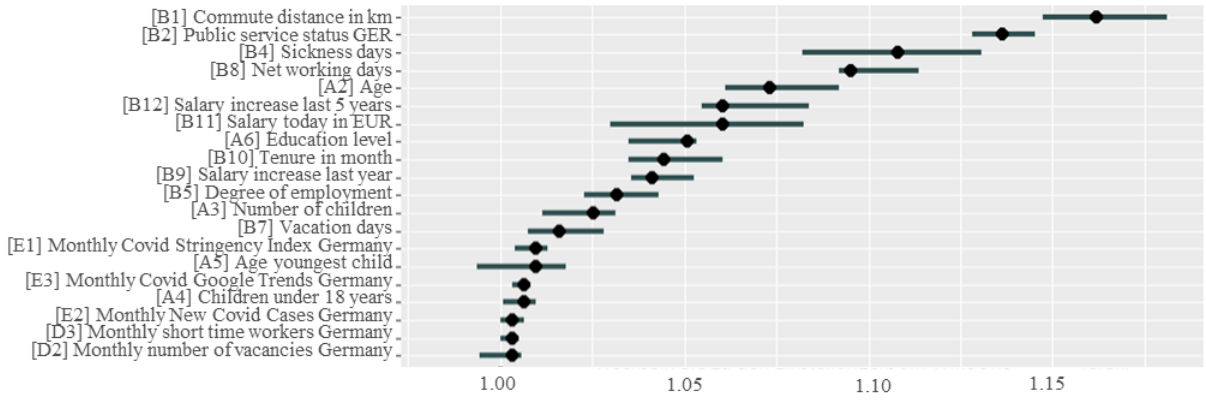


Figure 10: Permutated feature importance for the top 20 predictors (relative change in performance, confidence interval 0.95)

Figure 10 reveals that several demographic and work-specific predictors have the highest feature importance. *B1_commute_distance* and *B4_sickness_days* are among the top three features, consistent with other studies using data from the HRIS and finding them significant (Rombaut and Guerry 2021). External features generally have low feature importance but still contribute to predictive power. However, when comparing feature importance with other studies, it is important to note that results may vary depending on the features included in the dataset, the organisational context and changes over time (e.g., before and after Covid-19). A limitation of this high-level global feature importance method is that it is not suitable for examining whether the feature increases or decreases turnover risk.

Accumulated Local Effects

ALE describes how a single predictor influences prediction (strength, positive/negative contribution) on average, considering all employees in that local interval (Apley and Zhu 2020). Compared to alternative visualisation techniques on a global level (e.g. partial dependence plots), ALE is preferred when predictors correlate (Apley and Zhu 2020; Molnar 2022). *Figure 11* highlights the ALE plots for the top 12 predictors, sorted by descending importance.

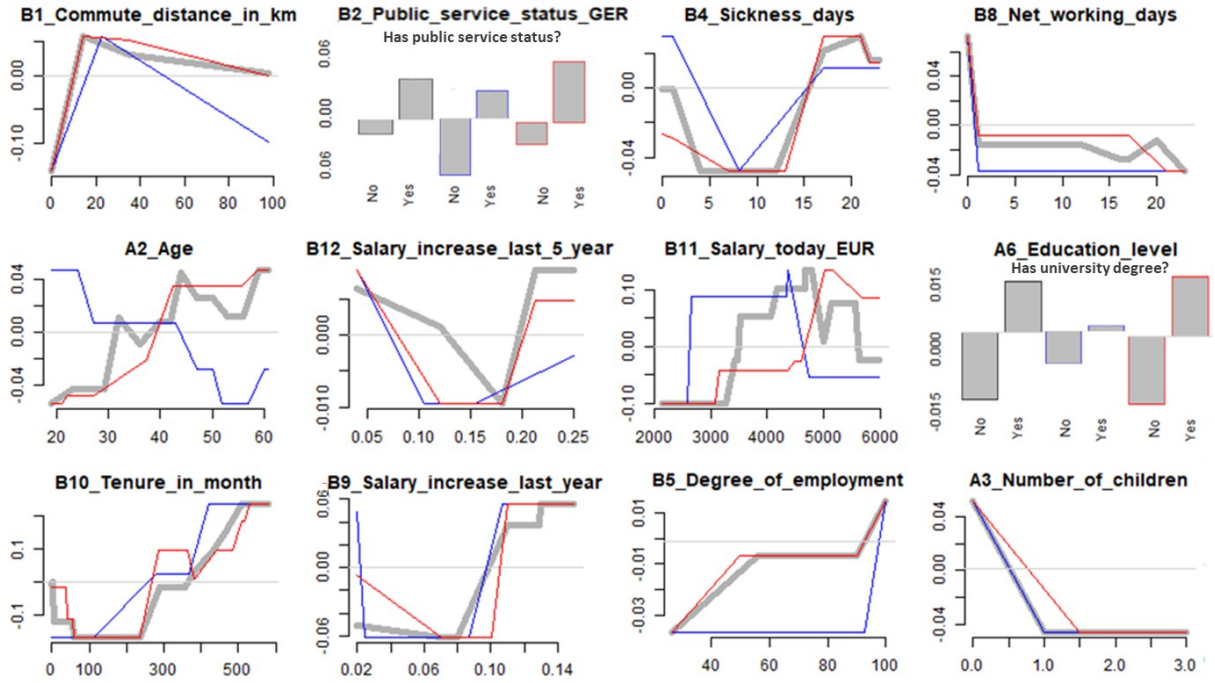


Figure 11: Accumulated local effect plots for the top 12 predictors according to permutated feature importance. Grey = all employees, blue = men, red = women.

The ALE representation helps compare results with existing empirical evidence. Consistent with research, we first find a higher turnover probability for employees with a higher *commuting distance* (Rombaut and Guerry 2021), but there seems to be a threshold around 15 kilometres – after which turnover probability no longer increases but slowly decreases. Second, we find that high – especially complete – absenteeism is an early indicator of turnover, as reflected in the *sickness days* and *net working days* ALE plots (Rubenstein et al. 2018, p. 42).

Additionally, we find that turnover decreases for men in line with age; interestingly, turnover among women increases. Since women form the majority in the considered organisation, this also determines the direction for all employees. This chart clearly shows heterogeneity between subgroups. One HR manager cited extensive measures to retain young mothers as a particular reason for this phenomenon. Also interesting is that there is no negative correlation between *length of service* and turnover, which is widespread in the literature, indicating organisational particularity. Instead, we see a U-shaped turnover probability in relation to tenure. Overall, the *age* and *tenure* results support the finding that these relationships are not as clear-cut in practice as is often assumed (Gray and Phillips 1994, p. 825).

Altogether, the ALE results are mostly in line with current findings; however, they also reveal organisational particularities, nonlinear relationships and interactions between predictors. However, one criticism of ALE is that it can lead to unstable and inaccurate predictions due to collinearity between predictors in intervals where instances are rare (Molnar 2022).

Local Feature Effects

Finally, we apply SHapley Additive exPlanations (SHAP), a local post-hoc explanatory method computing the effect of each predictor at the employee-specific level (Lundberg and Lee 2017). SHAP's local approach is similar to its popular alternative LIME, in that it divides the complex ML model into several mathematical vicinities (Chowdhury et al. 2022). What distinguishes SHAP explanations from LIME and others is their additive nature due to the game-theoretic approach, which facilitates a simpler understanding. SHAP breaks down the probability of voluntary turnover for each employee into SHAP values that quantify increasing or decreasing effects. Ultimately, the sum of the SHAP values is the difference taken from the average predicted probability of turnover for all employees (Lundberg and Lee 2017; Erel et al. 2021). *Figure 12* shows the relevant predictors with their values for two different employees, ordered by their additive explanatory contribution.

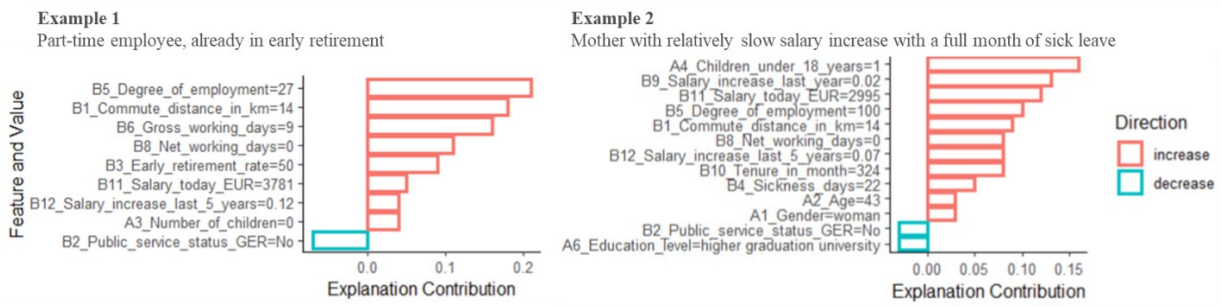


Figure 12: Top-10 SHAP values for two employees successfully predicted turnover candidates (true-positive)

In *Figure 12*, a part-time employee (*Degree of employment* 27%, *Gross working days* 9) did not attend work in the last month for unknown reasons (*Net working days* 0). Together with the existing *early retirement rate* (50%) and salary-related predictors, this indicates a higher turnover risk. Interestingly, in this particular case, the *Salary today in EUR* (3,781) and *Salary increase in the last 5 years* (12%) increase turnover probability, which is not clear in the ALE plots (Figure 11). Moreover, the explanatory contribution of *Tenure*, *Sickness days*, *Age* and *Education level* is relevant in example two but not in example one.

In summary, SHAP values at the individual (local) level for both examples are mostly consistent with ALE results. Nevertheless, a closer look at the local (employee-specific) compared to the global (company-wide) level offers more opportunities to challenge the results with existing evidence. The different selection of relevant predictors, as well as the contrary explanatory contribution, argues for interactions between predictors and different reasons for turnover in employee subgroups. Thus, our

results support the rationale for using ML and post-hoc methods to study employee turnover (Erel et al. 2021, p. 3247). As discussed later, it should be noted that local post-hoc explanatory methods like SHAP are often criticised because their explanations might be unstable and can therefore provide (intentionally) misleading explanations (Ghassemi et al. 2021; Vale et al. 2022).

Discussion

The increasing complexity of ML-based inductive research methods and algorithmic HRM leads to challenges, most notably ML opacity. Consequently, stakeholders may not act optimally, based on the ML predictions and insights proposed by algorithms, or accept the results (Meijerink et al. 2021, p. 2550). Similarly, ML-based inductive research offers a useful methodology only if the ML models are sufficiently transparent to extract insights.

Predictive performance vs. transparency trade-off

Accordingly, we come back to RQ1. Our results empirically demonstrate a significant trade-off between predictive performance and transparency in real-world employee turnover prediction data that is not found in artificially simulated datasets used in similar studies (Choudhury et al. 2021; Chowdhury et al. 2022). This suggests that artificially generated datasets may not holistically capture the sophistication of the various causes of employee turnover so that transparency-by-design approaches may be sufficient. Our results suggest that real-world HR prediction tasks like employee turnover prediction face a level of complexity that cannot be adequately solved by simpler transparency-by-design algorithms. The relationships between predictors and employee turnover elude linear relationships, making nonlinearities and heterogeneous interactions essential for accurately predicting the multifaceted causes of turnover (Putka et al. 2018, p. 721). Thus, the transparency-by-design approach may not be applicable in some real-world HR applications, leaving complex ML models. This supports recent theoretical research citing that opacity is a key characteristic and default of ML due to system-based opacity, caused by ML algorithm complexity and the scale required for meaningful application (Burrell 2016; Kellogg et al. 2020; Langer and König 2023).

Extracting multifaceted relationships

To understand the rationale behind complex ML model predictions, such as the random forest model, we propose three post-hoc explanatory methods that support the extraction of multifaceted insights on predictions. The three proposed XAI methods support translating the patterns used by opaque ML algorithms into human-understandable results. These patterns can also be multifaceted, such as nonlinearity and heterogeneous feature interactions. To accomplish this, all three proposed methods account for the extent to which predictors influence employee turnover prediction. However, each method takes a distinct mathematical approach and focuses on different aspects of information (e.g. local vs. global). ALE helps identify nonlinear and heterogeneous causes of turnover among employee

subgroups, including otherwise unnoticed insights; for example, in this empirical setting for turnover, (1) a U-shaped relationship with commuting distance (Figure 11), (2) interactions between the predictors age and gender (Figure 11), specifics of relevant predictors in diverse subgroups (Figure 12). SHAP highlights the specifics at the local level, allowing for the analysis of individuals. Such inductive findings can serve as the basis for theory development or complement deductive methods for deriving prescriptive and generalisable implications (Yuan et al. 2021, p. 3).

Criticism of post-hoc explanations

Finally, we discuss limitations of post-hoc explanations and their consequences for appropriate use. Other studies use post-hoc explanatory methods to demonstrate their capabilities for deeper investigation, although transparency-by-design approaches using linear logistic regression are sufficient (e.g., Choudhury et al. 2021; Chowdhury et al. 2022). In these cases, post-hoc explanatory methods (especially on local level) have been seriously criticised for high-stakes decision-making like HRM by technical experts, due to several risks resulting from the simplistic inaccuracies, and instead call for transparency-by-design approaches (Rudin 2019). Similarly, healthcare researchers refer to post-hoc explanations as a "false hope" because stakeholders may misinterpret the ML model's capabilities, mainly due to possible explanation inaccuracy (Ghassemi et al. 2021, 745) resulting from approximating the ML model to the real-world as well as the post-hoc explanatory method to the ML model. Due to this heuristic nature, the rationale behind predictions must be carefully questioned (Cheng and Hackett 2021). As a result of the resulting unreliability and superficial character of post-hoc explanation, Ghassemi et al. advise to not use post-hoc explanatory methods to access possible biases towards certain populations, to reassure for correct individual decisions, to increase trust or to justify acceptance of the ML model. They therefore call for using global explanatory methods only to understand how the ML model behaves at a global level and emphasise that they should be combined with rigorous prediction validation processes for different populations (Ghassemi et al. 2021, 745). Legal studies follow a similar line of reasoning, based on technical limitations such as results instability, stating that post-hoc explanatory methods cannot establish abstinence of discrimination caused by various biases embedded in ML models (Vale et al. 2022).

Justification for post-hoc explanatory methods

Concluding, in response to RQ2, post-hoc explanatory methods help examine the rationality of complex ML models to understand why they make incorrect predictions or reveal possible adverse impacts at the global level. However, given their limitations, they should be used with caution to justify individual-level personnel decisions. Post-hoc explanations should not be used as the only truth when formulating and justifying decisions where the stakes are high, but they are a helpful addition. Together, we argue for a nuanced perspective on the justified use of post-hoc explanations methods in circumstances where the transparency-by-design approach fails and managers are aware of its limitations. In these cases,

despite the criticisms, post-hoc explanatory methods are the most feasible way to mitigate the trade-off between predictive performance and transparency. Furthermore, as they provide only an approximation of complex ML models, they may serve as a source of information but cannot be regarded as the definitive ground truth. Thus, the validity of ML predictions and post-hoc explanations must be critically questioned following existing evidence, theory and domain knowledge to ensure causality (Erel et al. 2021, p. 3245). As an illustration, the plausibility of the relationships presented in the ALE plot (Figure 11) is supported by employee turnover studies indicating similar results. Similarly, the consistency of patterns discovered in relation to expertise can be evaluated against other organisation-specific information, such as survey-based data and exit interviews.

Implications for research

We contribute to the ML in HRM literature in three ways. First, we empirically demonstrate the trade-off between predictive performance and transparency in real-world data. In this context, our research documents that ML opacity may be unavoidable in some real-world HR prediction tasks, as the underlying algorithms must have considerable complexity to achieve adequate performance. Therefore, transparency-by-design approaches are not always applicable, as predictive performance does not sufficiently solve the prediction task. Second, we provide a nuanced perspective on the justified use of post-hoc explanatory methods, i.e. to mitigate the trade-off between predictive performance and transparency when it occurs in the HRM application – and transparency-by-design approaches are not sufficient. Third, we demonstrate the complementary use of local and global post-hoc explanatory methods to understand nonlinearities and heterogeneity in complex ML models.

Together, these three contributions have a methodological implication and can serve as a guideline when applying the ML-based inductive research method. Based on the extent of the trade-off with different algorithms during the experimentation phase, researchers may decide to adopt a transparency-by-design approach if it is suitable to solve the prediction task adequately. Alternatively, the three proposed post-hoc explanatory methods can help extract additional nonlinear and heterogenous insights. In contrast to existing studies that use a single post-hoc explanatory method (Choudhury et al. 2021, Chowdhury et al. 2022), we provide a broader picture of available technical solutions and demonstrate the joint use of local and global post-hoc methods in a complementary manner. However, for both methods, reliance on the human ability to make ethical and moral judgments, in order to critically question and rigorously validate the results of ML algorithms should be emphasised in both research and practice. To advance HRM knowledge, we therefore advocate a hybrid methodological approach, combining ML-based inductive, exploratory findings with validation by existing evidence and theories or by subsequent deductive, confirmatory studies. Accordingly, HRM research benefits from ML-based methods by (1) testing and refining theories and (2) expanding the explanatory range of theories (Leavitt et al. 2021). Additionally, we also make a secondary contribution to the HRM literature on employee turnover by

suggesting that examining nonlinearities and heterogeneity is important in fully capturing the intricacies of the various pathways resulting in turnover.

Implications for practice

For HRM practitioners, the well-known promises of objective and accurate ML-based decisions are only valid if models are not kept opaque. Instead, causality behind predictions must be verified through a hybrid human-ML development process before they can be used for individual decision-making (van den Broek et al. 2021). In this context, applying knowledge of the revealed trade-off between predictive performance and transparency guides a more informed algorithm selection in ML development. As technically-oriented functions such as data scientists may lack an understanding of opaque ML models' impact on HRM applications (Charlwood and Guenole 2022, p. 2), we suggest that organisations invest in educating HR staff about the consequences of ML complexity. In evaluating the potential impact on individual employees, HR managers can then weigh the predictive power or transparency of ML models on a continuum.

Educating HR decision-makers about the capabilities and limitations of post-hoc explanatory methods is also advisable; otherwise, they may not be able to 'use their tacit experience and social intelligence (based on intuitive thinking) to determine the accuracy of the model' when using complex ML models (Chowdhury et al. 2022, p. 20). Properly applied complex ML algorithms combined with post-hoc explanatory methods can mitigate the trade-off between predictive performance and transparency. This contributes to realise ML's potential to study and predict voluntary employee turnover and understand its multifaceted causes. The ALE plots' global explanations identify possible organisation-wide improvements or new retention strategies targeting influencing predictors, e.g. we find that commuting distance is an important turnover determinant, implying that a higher home-office ratio might be an effective countermeasure. ML predictions can be used in conjunction with nonlinear insights from ALE or individual-level explanations from SHAP to identify heterogeneous retention strategies for departments and demographic subgroups, or to personalise for employees, which would not be possible with linear models (Chowdhury et al. 2022, p. 15).

Ultimately, and particularly for public sector organisations, post-hoc explanatory methods might be a door-opener for ML-based methods. Public organisations must be highly transparent in their decision-making due to their high societal responsibility. Consequently, many public organisations have not yet started integrating ML-based solutions into their operations, one central reason for which is legal uncertainty related to the lack of transparency of such approaches. This was also one main reason for the examined federal agency in testing post-hoc explanatory methods and might be of similar interest for many other public sector executives seeking solutions in integrating reliable and trustworthy algorithmic HRM in their day-to-day business (Chowdhury et al. 2022, p. 24).

Limitations and directions for future research

Our first limitation is that inductive HRIS-oriented research does not provide deep insights in the same way that studies based on surveys do. Thus, pure data from HRIS and an inductive perspective should be used to complement existing theory-oriented research methods, e.g. for characterising risk groups or sub-organisational specifics (Rombaut and Guerry 2018, p. 97). In future research, the framework presented herein with post-hoc explanatory methods can be applied not only to data available in HRIS, but also to psychological studies' survey-based data. Thus, the inductive research method can serve as a complementary framework to study multifaceted relationships between multiple predictors. Unlike deductive research, comprehensive theory is not required to specify relationships in advance, thereby helping identify nonlinear and heterogenous relationships also for psychological constructs that predictors in studies based on linear models might miss (Putka et al. 2018, p. 690). Accordingly, we endorse the management literature (Leavitt et al. 2021; Valizade et al. 2024) by recommending ML as a powerful tool for quantitative research. Our results herein support the prioritisation of algorithms in terms of "transparency-by-design" or more complex algorithms in coordination with post-hoc explanatory methods.

Second, it is important to note that advantages and disadvantages of the three applied post-hoc explanatory methods may not be generalisable to options beyond the scope of this work. Further, HRM research should be aware of the rapid development of ML algorithms and XAI, as computer scientists attempt to resolve the trade-off by developing new transparency-by-design algorithms to increase their predictive performance, as well as new post-hoc explanatory methods (e.g., Arrieta et al. 2020).

Third, we focus on one federal agency in an in-depth examination, which means we forfeit some generalisability (Yin 2013, p. 325). We encourage future research to investigate implications of the trade-off in diverse other (multi-)national settings and in other ML-based prediction tasks in HR besides employee turnover prediction, such as employee selection, training and management.

Conclusion

This study discloses the trade-off between predictive performance and transparency in ML empirically demonstrated in a real-world HRM application, meaning that complex – and therefore opaque – algorithms have significantly better predictive performance. For sophisticated prediction tasks such as employee turnover, the underlying algorithms must have considerable complexity and therefore cannot be adequately solved by simpler transparency-by-design ML algorithms. However, by applying three post-hoc explanatory methods to successful but opaque ML models, insights are gained that include nonlinear and heterogeneous causes of employee turnover. We argue that post-hoc explanatory methods help mitigate the trade-off if they are used properly according to their limitations and are not blindly trusted, which is why we emphasise a nuanced perspective to their justified use. We hope that this paper

motivates further research regarding ML transparency as a necessity to pave the way for an ethically and a legally compliant ML-augmented decision-making process that benefits the organisation and – most importantly – all employees.

2.2 Study 2 | Towards fair Human Resource Analytics: Introducing a sociotechnical framework

Abstract

ML algorithms have the potential to revolutionise human resource management (HRM) by providing unprecedented opportunities for data-driven decision-making, but they also pose challenges such as algorithmic fairness (AF). Several automated bias mitigation approaches have been proposed to achieve AF by reducing the amplification and prevalence of discrimination through statistical measure optimisation. However, the social sciences emphasise that automated approaches lack the necessary background and context to reflect the human values relevant to HR Analytics. Building on sociotechnical systems theory, and applying a design science methodology, we introduce a framework for AF assessment that emphasises the procedural intertwining of partial automation and human augmentation. Our findings contribute to understanding how a socio-technical system design bridges the gap between the disciplines of IS and HRM. The proposed framework guides interdisciplinary research on non-discriminatory findings and provides guidelines for practitioners to address the challenges of AF in HR Analytics.

Introduction

With the proliferation of ML over the last decade, algorithms have emerged as a powerful tool in HR, offering unprecedented opportunities through the application of business analytics for data-driven decision-making and insight generation. This study examines AF assessment in HR Analytics, which is defined as a “practice enabled by information technology that uses descriptive, visual, and statistical analyses of data related to HR processes, human capital, organisational performance, and external economic benchmarks to establish business impact and enable data-driven decision-making” (Marler and Boudreau 2017, p. 15). In short, HR Analytics is the application of business analytics in HR to improve decisions and/or automate HRM activities, resulting in increased efficiency and effectiveness (Meijerink et al. 2021). Contrary to its great potential, the application of HR Analytics in high-stakes decision-making, which can significantly impact individuals' lives and careers, has raised concerns, particularly because ML models can incorporate non-traceable and unfair biases (Cheng and Hackett 2021). Practical examples have shown that ML-based HR Analytics systems can inadvertently lead to bias against protected demographic groups or minorities. For example, Amazon stopped developing an ML-based recruitment system because it discriminated against women (Meyer 2018). In addition, Facebook's ad-targeting algorithms disadvantaged women in job postings because they were less likely to see ads from companies hiring predominantly male employees (Teodorescu et al. 2021). In the background, decision-makers unfamiliar with the capabilities of algorithms may be deceived or misled into making incorrect decisions, or they might even use the algorithms as an excuse to make discriminatory decisions (O'Neil 2016). This "dark side" of analytics is a growing concern that generates several emerging research questions (Mikalef et al. 2022), e.g., What processes need to be established to minimise bias in ML applications?

The HRM literature has already recognised that bias is not inevitable, and the discipline has moved towards applying available tools and approaches (Charlwood and Guenole 2022) provided by IS research to address possible unfairness (for a review see for example Mehrabi et al. 2019). In this context, two distinct research strands regarding AF assessment have emerged, focusing either on enhancing human-centred augmentation and ML explainability (e.g., Arrieta et al. 2020; Chowdhury et al. 2022) or statistical fairness measure optimisation (e.g., Speer 2021; Rottman et al. 2023). The literature on ML explainability generally proposes an augmentation approach in which fairness can be ensured by better tracing how and why an ML model makes predictions (human-in-the-loop). In contrast, the literature on AF defines (partial-) automated fairness measure optimisation approaches with the goal to quantitatively reduce differences between protected groups such as gender, nationality or age. Unfortunately, both approaches have largely remained separate and thus overlooked the critical interplay between automated bias mitigation approaches and explainability-enabled human-

augmentation in AF assessment. Motivated by this gap, this study aims to investigate how the two separate approaches can interactively mitigate discriminatory biases to promote AF in HR Analytics.

To fulfil this research goal, it is necessary to optimise a holistic decision-making system that includes mutual interactions between people, organisational structures and technology to find solutions to potential unfairness in HR Analytics. This is why current IS research stresses the application of sociotechnical system theory as a holistic perspective (Dolata et al. 2022; Kordzadeh and Ghasemaghaei 2022). In this study, a socio-technical framework (dimensions: interface, subprocesses and interconnections of processes) is proposed that combines state-of-the-art automated bias mitigation approaches with human augmentation through an interface relying on improved ML explainability through XAI methods. We empirically demonstrate the utility of this framework in a complex, real-world AF assessment scenario by identifying nonlinear predictor effects among multiple protected groups from ML models, in order to predict employee turnover in a public-sector organisation. We contribute to the literature by answering research questions from the interdisciplinary research streams and by bridging a gap between them in three ways. First, we respond to the call for design science research to innovatively design, construct, analyse and evaluate artefacts that enable algorithmic system developers to implement domain-specific dependencies through design approaches (Holstein et al. 2019; Kordzadeh and Ghasemaghaei 2022). Second, we directly respond to Rottman et al. call for advances in the field of XAI to be applied in combination with modern automated bias mitigation approaches (Rottman et al. 2023). Third, we contribute to the open research questions on how to establish adequate control mechanisms and processes in joint human-algorithm decision-making, as well as stakeholder involvement in AF assessments (Dolata et al. 2022)

In the following section 2, we review the relevant literature on AF from different perspectives, starting with the technical perspective, moving to the social perspective and then examining how a sociotechnical systems perspective can overcome mutual limitations. In Section 3, we briefly describe the design research method, followed by a detailed description of the proposed framework (Figure 14). Section 4 demonstrates the utility of the framework in a field setting. Section 5 provides a discussion as well as implications for research and practice. Finally, section 6 concludes the paper.

Literature review

Common understanding of Algorithmic Fairness (AF)

In order to study AF from an interdisciplinary perspective, we first have to develop a common understanding thereof among different research fields. A popular definition of AF is “the absence of any prejudice or favouritism towards an individual or a group based on their inherent or acquired characteristics” (Mehrabi et al. 2019, p. 1). Related to this notion, the literature refers to adverse impacts as differences between majority and minority groups in employment-related opportunities, e.g., hiring,

promotion and termination, distributed through the use of algorithms (Charlwood and Guenole 2022). Adverse impacts occur when seemingly neutral models have discriminatory effects on protected group variables such as race, gender, religion, sexual orientation, ethnicity and others. Even if models do not explicitly include protected group information as a predictor, they can still unintentionally result in adverse impacts. Predictors used in employee turnover prediction models can include demographic data such as a zip code, which correlates with ethnicity (Castille and Castille 2019; Speer 2021).

To achieve AF, non-discrimination is a fundamental requirement relevant to several regulations. The European General Data Protection Regulation states that in a minimal interpretation algorithms do not explicitly include sensitive data or protected class variables; otherwise, the model is referred to as 'disparate treatment' and thus can be subject to legal scrutiny. The maximum interpretation states that no predictors can be included that correlate with the above protected class variables and may thus lead to an adverse impact (Goodman and Flaxman 2017). Likewise, HR Analytics on an individual employee level may be subject to civil rights laws in the United States, which protect individuals from discrimination based on protected class variables (Castille and Castille 2019). According to the Civil Rights Act, adverse impact ratios are calculated with the goal of evaluating whether unjustifiable biases in the ML model lead to different selection ratios in terms of positively classified candidates: the adverse impact ratio must not violate the "4/5 rule," i.e., when the selection rate of a group is less than 80% of the group with the highest selection rate (Civil Rights Act 1964). However, only when these differences cannot be justified or explained by real-world differences is it considered illegal discrimination (Mehrabi et al. 2019). For example, adverse impacts based on job-related predictors are not unlawful under the U.S. courts' interpretation, as long as predictive accuracy is similar for all groups and no alternative predictor with comparable effectiveness and fewer adverse impacts is available (Charlwood and Guenole 2022).

Technical approaches to AF

The IS literature proposes several approaches to automatically eliminate most (up to a certain threshold) or all adverse impacts between protected classes, based on a single statistical fairness metric. Technical approaches to AF thus aim to "detect, quantify, and subsequently mitigate disparate harm (or benefits) across subgroups affected by automated decision-making" (Dolata et al. 2022). They are usually benchmarked on their capabilities to optimise the trade-off between statistical fairness measures and predictive accuracy, also known as the "diversity-validity dilemma" (Rottman et al. 2023). A detailed review of these technical approaches is beyond the scope of this paper, so we refer the reader instead to Mehrabi et al. (2019). More recently, HRM researchers have recognised that achieving AF is becoming increasingly complex, as it is difficult to distinguish meaningful patterns from unwarranted patterns related to adverse impacts, as the amount of data increases and the underlying ML models become more sophisticated. Therefore, the various options developed in IS research are a valuable addition to existing

approaches from the social science perspective (e.g., Rottman et al. 2023). These automated bias mitigation approaches can be separated into three categories (Mehrabi et al. 2019):

- Pre-processing approaches, which transform input data for training the model.
- In-processing approaches, which change modern ML algorithms to remove bias during model training, either by incorporating changes into the objective function or imposing a constraint.
- Post-processing, which is performed after training by accessing a holdout (not used for training) dataset.

Furthermore, technical approaches can be divided into 'group-agnostic', by applying an identical logic to all groups equally, and 'group-specific', applying different sets of logic to different groups. Rottman et al. argue that the field of ML may not know that while group-specific approaches can reduce adverse impacts while maintaining validity, these strategies explicitly treat each group differently (Rottman et al. 2023). Moreover, since group-specific outcomes depend on group membership, and groups are not subject to the same standards, group-specific approaches represent unequal treatment when used in HR Analytics (Civil Rights Act of 1964). This paper follows this line of reasoning and therefore includes only group-agnostic automated bias mitigation approaches.

Limitations of the technical approaches to AF

The primary constraint in the technical approaches to AF is the notion of fairness, which is reduced to a single mathematical expression (differences between protected groups, differences compared to similarities between individuals, etc.) based on the assumption that information about the biases to be addressed is known a priori (Dolata et al. 2022). As a result, in the IS literature, the terms 'adverse impact' and 'bias' are often used interchangeably. Conversely, in the social science literature, the term 'bias' is only used when the adverse impact is not justified by real-world phenomena (Charlwood and Guenole 2022). If it is not always desirable to eliminate all adverse impacts, then the optimisation approach of the validity-diversity dilemma loses its legitimacy. At worst, correcting so-called 'explainable' adverse impacts that are not based on bias against protected groups but are justified by real phenomena might result in "reverse discrimination" (Mehrabi et al. 2019). However, even if adverse impacts do not shift toward reverse discrimination, model validity is affected by removing group differences caused by explainable differences (Charlwood and Guenole 2022). Incorporating this basic distinction into automated bias mitigation approaches is particularly challenging.

Another aspect where human augmentation is indispensable is the selection – from several options – of suitable fairness measures, i.e., various quantitative approaches to establishing a concept of fairness, such as the popular alternatives of demographic parity, equality of opportunity and equality of opportunity (Mehrabi et al. 2019). This is necessary because multiple statistical fairness measures or adverse impact constraints are inherently incompatible, and combinations of fairness measures are even mathematically impossible according to the impossibility theorem (Teodorescu et al. 2021). Therefore, depending on the HRM data generating processes and context, domain experts must decide which

fairness definitions should take precedence. Finally, Speer notes that it is often non-trivial to determine which outcome of an ML model leads to a disadvantage. For example, in a classification model that predicts employee turnover, either the positive or the negative prediction might be favourably affected (Speer 2021). If the company intends to take retention measures (e.g., a pay raise) for those employees with higher turnover probabilities in the positive class, they will be supported. If the company seeks to promote those who are less likely to leave the company, employees with lower turnover probabilities in the negative class will be favoured. In HR Analytics, HRM professionals deem that this decision be neither automated nor delegated to technical staff developing an ML model (Speer 2021).

In summary, automated bias mitigation has limitations in terms of (1) the distinction between justified explainable group differences and discriminatory bias, (2) selecting and prioritising appropriate fairness measures from multiple incompatible alternatives and (3) deciding which protected group is expected to experience negative consequences. Due to these limitations, it is clear that AF assessments cannot be fully automated due to the social, behavioural and organisational aspects of AF that are of great importance for decision-making in HR Analytics (Kordzadeh and Ghasemaghaei 2022).

Sociotechnical approaches to AF

Despite the above limitations, technical automated AF approaches may be essential to mitigate bias as the amount of data and complexity of algorithms increase and humans alone may be unaware of the myriad ways in which algorithms can be unfair (Teodorescu et al. 2021). In social science research, sophisticated automated bias mitigation is increasingly recognised as opening up new and innovative ways to address the dilemma between diversity and validity (Rottman et al. 2023). Ultimately, a purely social approach may be inappropriate if it does not consider how the proposed social solutions can be operationalised, given the multiple technical options and algorithms available (Dolata et al. 2022). For this reason, we follow current IS research (Dolata et al. 2022; Kordzadeh and Ghasemaghaei 2022) that theorises AF as a sociotechnical construct. According to the sociotechnical system theory, AF depends on the mutual influence between technical and social structures, as well as between instrumental and humanist values, to achieve a state of coherence (Sarker et al. 2019). This state of coherence optimises the AF of overall system outputs rather than the adaptation of the structure or internal processes of a single component – as traditional IS research seeks to do for the technical component. This is an important point because only a sociotechnical approach recognises the individuals directly affected by decisions, including their overarching consequences on the state of the overall system, when feedback loops cause varying notions of AF and their appropriate operationalisation (Kordzadeh and Ghasemaghaei 2022).

A sociotechnical system for AF can be divided into two subtasks: (1) predicting the likelihood of certain outcomes (predictive accuracy, adverse impact ratios, etc.) based on relationships modelled from data and (2) judgement of trade-offs in situations that cannot be operationalised with a measurable fairness

metric (Teodorescu et al. 2021). ML outperforms humans in predictive tasks, especially in scenarios characterised by high complexity, due to their ability to analyse large datasets efficiently. Conversely, humans outperform ML in terms of exercising judgment by understanding the implications of different predictive outcomes in a broader context (Agrawal et al. 2018). This means that in a sociotechnical approach to AF, algorithms and humans should take on different tasks (Raisch and Krakowski 2021). Eventually, in a sociotechnical methodology, humans and ML need to complement each other by relying on their strengths and overcoming each other's weaknesses – as displayed in the typology of how human-augmented approaches to AF (Figure 13) can be realised (Teodorescu et al. 2021, p. 1489).

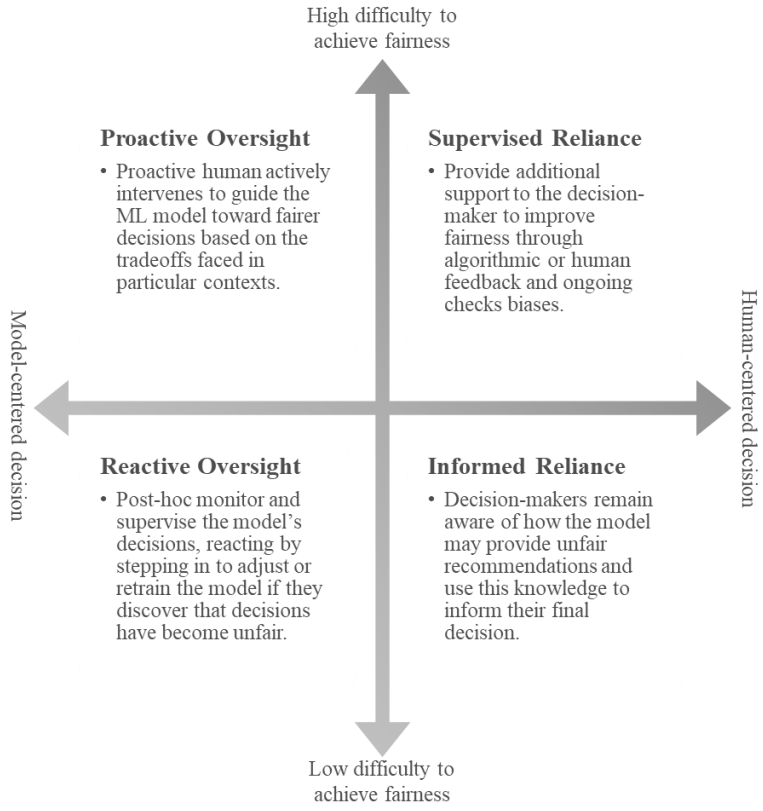


Figure 13: Typology of human-augmented AF assessment, adapted from Teodorescu et al. 2021

ML explainability enables sociotechnical approaches to AF

The literature on HR Analytics emphasises that ML explainability is a basic prerequisite for AF, since opacity impairs the ability to check predictions against the HRM professional's intuition (Chowdhury et al. 2022), existing evidence (Charlwood and Guenole 2022; Meijerink et al. 2021), critical thinking about the logical background of predictions, causal inferences and subgroup differences (Putka et al. 2018). Thereby, ML opacity impedes this co-evolutionary hybrid human-ML augmented learning process that eventually leads to a highly accurate, fair and unbiased ML model (Raisch and Krakowski 2021). Likewise, the technical literature on ML argues that explainability is closely linked to fairness (Zhou et al. 2022), since it is essential to identify possible predictors leading to unfair bias (Lee 2018).

This is why the IS literature has developed numerous options to achieve higher ML explainability while (largely) preserving the model's validity and accuracy (Arrieta et al. 2020). Likewise, the IS literature aims to understand how these XAI methods can be embedded and enhance AF. Dodge et al. (2019) explore why the need for enabling human augmentation in AF assessment requires different styles of explanation (e.g., local vs. global). They find that global explanations can be used to enable humans to assess biases in the model in general, and that local explanations can be used to examine individual AF. These findings underline the narrative of the need for a sociotechnical approach to AF. Binns et al. (2018) found that the case-based explanation style (i.e., providing examples to justify decisions) had the most significant impact on perceived justice associated with an algorithmic decision, whereas Dodge et al. (2019) reported that sensitivity-based explanations (i.e., amount of change in a variable needed for the decision to change accordingly) were more effective than case-based explanations at making biases transparent. Haque et al. (2023) suggested using easy-to-understand XAI explanations and visualisations in HRM such as predictor effects. In addition to the design of transparent ML models through XAI, an interface is needed for augmentation with humans in the loop during evaluations when adverse impacts can be justified (Zhou et al. 2022). Simply put, an interface needs to be able to connect ML models' predictions and human judgement in such a way that humans have the competencies and cognitive abilities necessary to understand the content (Bader and Kaiser 2019).

To summarise, while social and technical research streams agree on the importance of explaining ML and the appropriate design of interfaces for socio-technical approaches to AF, there is still little evidence on the selection of XAI methods and their combination with automated technical approaches to AF.

Knowledge gap

While research has addressed the relationship between the social and technical perspectives on AF, they rarely go beyond identifying problems in each area, so a holistic and comprehensive framework based on sociotechnical theory remains absent (Dolata et al. 2022). For instance, research on the human-in-the-loop claims that introducing human control or a last word will lead (to a certain extent) to AF, but a sociotechnical perspective reveals that this idea of a righteous and critical person independent from an algorithms is problematic (Dolata et al. 2022). Instead, an algorithm empowers and constrains humans through the specific information provided. It is therefore important to understand how various information, such as prediction itself, fairness measures and explanations provided by XAI (such as predictor effects), can be simultaneously integrated into human assessment to achieve AF. Here, the proposed related frameworks remain highly normative in terms of how automated information generation design and augmented assessment processes need to be interlinked (e.g., Teodorescu et al. 2021). Moreover, analytics practitioners, including ML developers and product managers, call for technical, procedural and managerial support to help address AF (Holstein et al. 2019; Kordzadeh and Ghasemaghaei 2022). Thus, the frameworks need to be expanded to capture complex mutual influences

between the technical and social parts of the ensemble (Dolata et al. 2022). In the following, we propose artefacts (processes and methods) based on existing technological building blocks of different IS research domains (human-machine collaboration, XAI, business analytics) that are integrated into a defined technical framework adapted to the specific requirements of HR Analytics.

Method and framework development

Due to the innovative and technical nature of the knowledge gap, this study follows a design science methodology. In design science, researchers seek innovative solutions to real-world problems and derive more general theories and insights from practice-oriented results (Gregor and Hevner 2013). Its goal is to create innovative artefacts that emerge from alternative solutions to significant real-world problems, and these may include models (abstractions or representations), methods (frameworks, algorithms or practices) and instances (processes or prototype systems). Developing artefacts contributes new knowledge to nascent theories in the form of operational principles or solutions (Gregor and Hevner 2013).

This section describes the proposed framework, which is shown in a schematic overview in Figure 14. In general, the sociotechnical framework for AF includes (1) an interface with the specification of information for human augmentation, (2) a technical subprocess (ML model training using automated fairness measure optimisation approaches) and (3) a social subprocess (ethical considerations and human augmentation) along with the interconnections between these subprocesses. Before discussing these artefacts individually, we should take note of three general key points when using the framework. First, the overall optimisation objective of the sociotechnical system output is the AF of an HR Analytics decision-making process (e.g., classification in recruiting, development or retention). In addition to quantitative measurement (e.g. demographic parity), which is based on historical predictions of decisions, AF can also be measured qualitatively (e.g. perceived fairness) during the audit (Newman et al. 2020; Langer et al. 2023). This audit determines whether (1) an algorithmic recommendation is approved and made available for practical decision-making, (2) iterative adaptation processes for ML models are required or (3) the achievement of unfair models is determined to be unachievable. Second, the specific design of individual components with multiple choices (e.g., fairness measures, XAI methods, AF pre-/in-/post-processing options) listed in the framework should not be understood as a ‘must’ and definitive list but rather as possible recommendations or options for sociotechnical systems design. Third, the four human-augmented approaches to AF (already depicted in Figure 13) proposed by Teodorescu et al. (2021), along with an ontology of when to use which approach, are important elements in this regard. In this framework, we embed these options as interconnections that can bridge the gap between social and technical subprocesses. Rather than determining the final location of the decision, one or more interconnections can be chosen to organise the procedural dependencies and freedom of business analytics/data science and HRM teams.

Interface

The interface is the central and comprehensive artefact that passes information from technical subprocess (e.g., ML model options and their predictions) to the audit, which is the final step in the social subprocess for human augmentation, including the final judgment. In the following, we discuss four essential information components that can be extended if required. The components are local as well as global post-hoc XAI methods, as they contribute to different notions of fairness (Dodge et al. 2019) and act as a monitor for the output history (Teodorescu et. al 2021) and visualisation of a model benchmark. Thereafter, we discuss the technical subprocess for creating these components.

A | Global explanations (e.g., predictor effects on groups)

Global explanations on a group level, such as predictor effects or predictor importance, help access bias on a (semi)-aggregated level for an organisation. In the case of fewer complex algorithmic options with sufficient accuracy, which is in itself interpretable (like linear regression or decision trees), these visualisations can be directly derived from the ML model. For complex ML models that remain opaque, post-hoc XAI methods should be used. The former option is preferable, as it loses less information due to the approximation applied in post-hoc XAI methods, albeit this is not always possible in highly sophisticated ML applications (Rudin 2019). For example, global post-hoc XAI method effects such as ALEs extract non-linear predictor effects on cohorts (Apley and Zhu 2020) and thus provide the necessary transparency to question rationality.

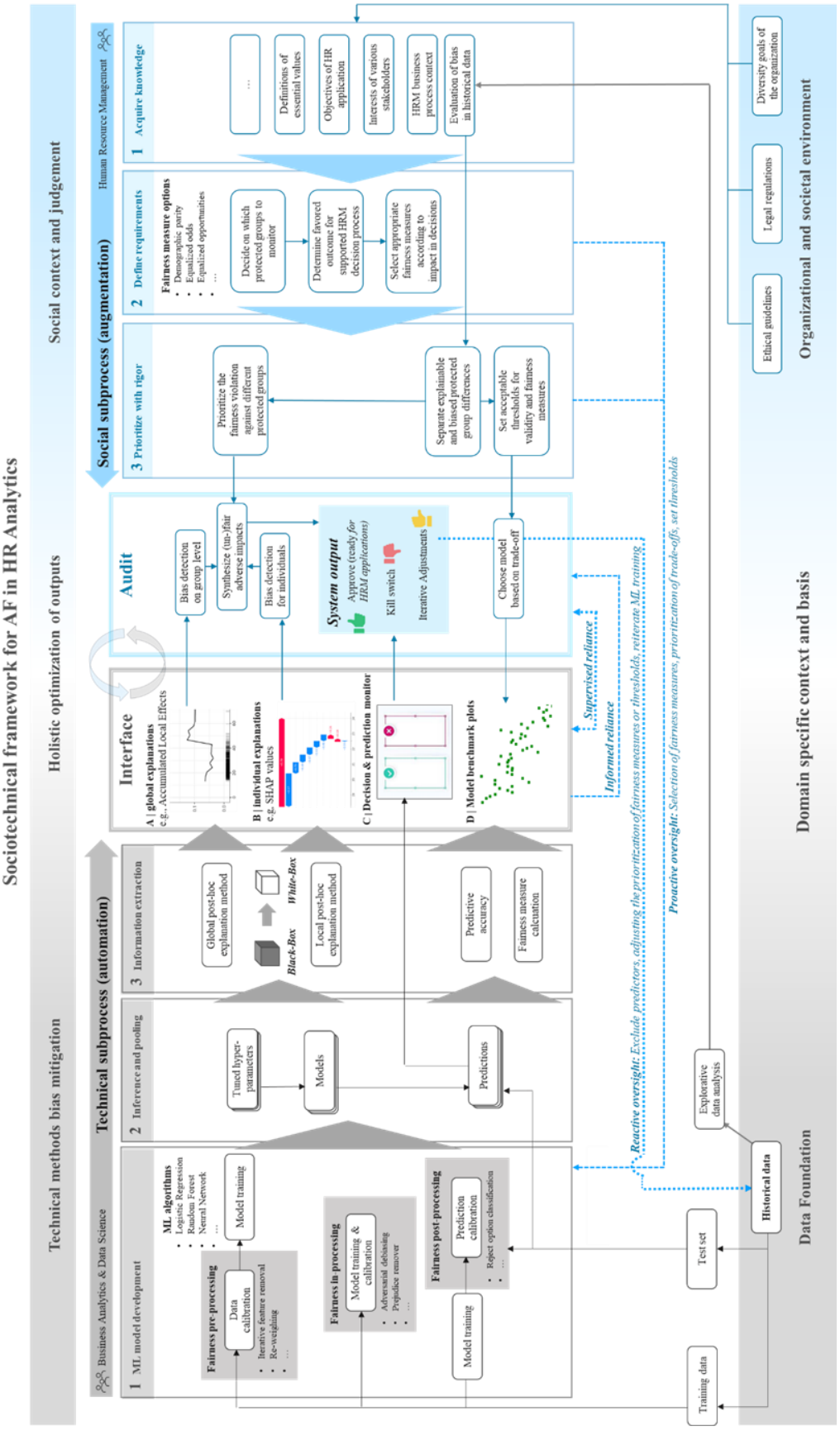


Figure 14: Proposed sociotechnical framework for an AF assessment in HR Analytics

B | Local explanations (e.g., predictor effects on individuals)

Identifying individual-level bias requires a deeper understanding of which factors and characteristics enable an employee-specific prediction. Local post-hoc XAI methods such as SHAP values are one example what can be used for this purpose (Lundberg and Lee 2017). It should be noted that the advantage of zooming in on specific environments with local post-hoc XAI methods also presents challenges in terms of stability, validity and ambiguity (Molnar 2022). Thus, specific technical knowledge in the implementation phase must be available to avoid pitfalls. Therefore, if ML explainability (especially at the individual level) is achieved through post-hoc XAI methods, care must be taken to provide clear, trustworthy and verifiable indications of predictor effects (Rudin et al. 2020). For this reason, the legitimacy of the information included should be considered in the context of the other components in the interface (Vale et al. 2022), even though some research suggests that AFs may rely solely on local post-hoc XAI declarations (Chowdhury et al. 2022).

C | Monitor historical decisions and predictions

The monitor tracks both the algorithmic advice proposed by ML models and the adjustments and final decisions made by humans. These historical records can be used to track over time the optimisation success of results from the holistic socio-technical system. Thus, the monitor could extend the traditional idea of human-in-the-loop, where humans and machines monitor each other from different perspectives and aggregation levels (supervised reliance) (Teodorescu et al. 2021). The perspectives may include aggregate AF measures, adverse impact analysis across multiple protected groups, confusion matrices and others.

D | Model benchmark plots on the trade-off between diversity and validity

Even if the automation process can filter out most suboptimal models, due to higher negative effects and/or lower validity compared to other options, the number of resulting non-trivially worse options may still be high due to the convex trade-off between diversity and validity (Arrieta et al. 2020). Therefore, some form of visualisation, such as a graph with axes for diversity (e.g., measured by average negative impact for protected groups) and validity (e.g., measured by accuracy), is an essential basis on which to allow people within the audit to decide which model to use.

Technical model development subprocess

In the technical subprocess, various options for ML model training are applied under the responsibility of business analytics or data science teams. The input for this process is historical data and information, provided in the form of a reactive or a proactive oversight (which predictors to exclude, initialisation of retraining, prioritisation of the trade-off between diversity and validity, acceptable thresholds for fairness measures, etc.). The technical subprocess includes embedded resampling strategies to achieve out-of-sample validation (training/test split). Proposed options are aligned with domain-oriented works in HR Analytics that present frameworks and proposals for ML development introducing the principles of

modern predictive models, common algorithms and post-hoc XAI methods (Putka et al. 2018; Chowdhury et al. 2022). It should be noted that each algorithm chosen as part of the 'model training' can be combined with multiple options for automated bias mitigation approaches. Since this step is fully automated, a considerable number of options (different AF optimisation approaches *x* number of ML algorithms *x* respective number of tested hyperparameter constellations) can be proposed during model development. The automated technical subprocess is thus able to provide many models with different prediction accuracies and adverse impacts resulting from the non-parametric and non-linear nature of ML algorithms (Arrieta et al. 2020).

The framework leaves flexibility in terms of selecting pre-, in- or post-processing options of automated fairness optimisation. However, it should be noted that group-specific automated approaches employed to reduce bias due to the HR Analytics domain-specific AF assessments are not shown in the framework because they treat protected groups differently (Rottman et al. 2023). Each of the models trained in automated model development is applied to test data in the inference step and registered (together with its predictions) on a database (pooling). In addition to the basic statistics (fairness measures, prediction accuracy) extracted during model development, relevant information about the ML model' rational to form predictions is displayed in the interface (e.g., XAI visualisations). The technical subprocess is triggered iteratively by the social subprocess using *reactive oversight* and *proactive oversight*. *Proactive oversight* provides the initial conditions (selection of fairness measures, prioritisation of the specification of the trade-off between diversity and validity, definition of acceptable thresholds, etc.) for the first automated model development iteration. In *reactive oversight*, an audit's decision to make iterative adjustments is made with respect to excluding spurious predictors that are likely to introduce bias from the historical data, or with respect to prioritising fairness measures or acceptable thresholds to repeat model training.

Social human-augmented subprocess

In social subprocess under the responsibility of the HRM team, the first step is to acquire the required knowledge from the societal context (ethical guidelines, legal regulations, etc.), together with the organisation's diversity goals. Specifically for the HR Analytics implementation, the general objectives and priorities of various stakeholders, processes in a specific HR Analytics implementation and a definition of essential human values form the knowledge base for defining requirements. In addition, exploratory data analysis can uncover biases in historical data resulting from the state of the art, which is an important source of information when seeking to uncover data-related biases before embedding them in complex ML models. The second step is to define and agree the details for the quantitative measurement of AF. This includes deciding which protected groups should be monitored (gender, nationality, etc.), determining the preferred outcome for the supported HRM decision-making process and selecting one or more appropriate fairness measures (e.g., according to the consequences of false-

positive and false-negative predictions). Third, once the requirements have been defined, the outcomes of the sociotechnical system must be prioritised with rigour (balancing diversity and validity trade-off, which violation to be addressed versus which protected groups, etc.). Here, an central step in the proposed socio-technical framework which represents a significant distinction to automated AF approaches is the separation of explainable and unfair bias from adverse impacts (Charlwood and Guenole 2022). If information required to separate explainable and biased protected groups is embedded in thresholds and passed to the technical subprocess via proactive oversight, the technical team can implement models that eliminate only harmful biases instead of all protected group differences.

The central decision locus is the audit, in which the results of the social sub-process are closely interlinked with the results of the technical sub-process represented in the interface. In the audit, primary social objectives are iteratively compared with details of the technical model options provided in the interface components. First, a model must be selected from several options, using a model benchmark plot. Then, unfair biases at the individual and group level are uncovered in the review, using the XAI visualisations for that specific model. This knowledge of the characteristics used by the ML model for the prediction contributes to a final summarised assessment of AF. Before a final decision is made, a safeguard mechanism, namely a check against historical decisions and predictions, is used. This check against ground truth ensures that the usual pitfalls of XAI methods, such as possible inaccuracy and instability, are mitigated (Vale et al. 2022). When a desired output of the sociotechnical system for the AF cannot be reached, another possible outcome of the audit involves terminating the ML models by using a 'kill switch'. Ultimately, this kill switch may be used to cancel the complete ML implementation project if the audit concludes that even when applying a sociotechnical approach for AF, the potential consequences and risks of unfairness outweigh the benefits (for a practical example see Meyer 2018).

As outlined in sociotechnical systems theory, a state of coherence in overall output should be achieved (Dolata et al. 2022). However, this requires monitoring and feedback loops when data, algorithms or human perceptions change, thereby implying a continuous need for audits (*informed reliance*). Alternatively, in a *supervised reliance* approach, the interface not only provides recommendations, but also includes ongoing checks on the quality of the final sociotechnical decisions. For example, integrated monitoring of historical ML predictions and human decisions can enable the analysis of when a human correction leads to potentially problematic higher adverse impacts for multiple protected groups (e.g., gender differences decrease but nationality differences increase). In this way, humans can continuously learn from ML monitoring. Similarly, a monitor could set up a feedback system between humans so that humans monitor others' decisions against model predictions. In complex decision scenarios involving humans (e.g., multiple fairness measures, multiple protected groups, sophisticated ML models), this allows for the most comprehensive guidance and promises an optimal combination of algorithmic and human advantages (Teodorescu et al. 2021).

Empirical demonstration

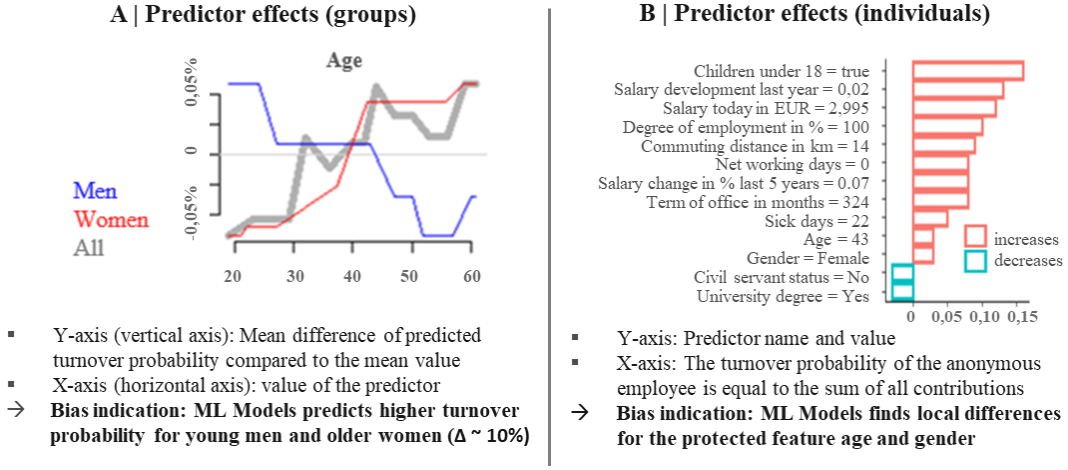
This section demonstrates the application of the framework, using excerpts from an ML model implementation process for employee turnover prediction in a German federal agency with more than 20,000 employees between 2021 and 2023. The model predicts individual voluntary (age-related causes are excluded) turnover probability within the next 6 months for each employee, using ML algorithms and a dataset with 30 predictors acquired over a period of 36 months. A complete presentation of multiple automated approaches in the technical subprocess, as well as multiple relevant ethical considerations in the social subprocess, is beyond the scope of the paper. Instead, we focus on demonstrating how the sociotechnical framework leads to a holistic optimisation in complex AF scenarios by iteratively combining technical and social subprocesses. Due to the extensive IS literature on the technical sub-process, we aim to focus herein on an automatic approach to reduce bias, namely the iterative removal of predictors, which we choose for several reasons: (1) It can be combined with any ML algorithm and prediction task (regression, classification, etc.), (2) it reflects the basic idea of the framework well, as it provides a human interface in each iterative step, (3) it makes it possible to use customer-specific rule sets (Dodge et al. 2019) and (4) it is already established in the HR Analytics literature (Speer 2021; Rottman et al. 2023). Likewise, demographic parity is used here and in this literature because it is easy to understand.

Social context and initial model training (proactive oversight)

Before technical training for the first models, HRM managers at the federal agency identified the protected group variables age (above average old, below average young) and gender as relevant for monitoring in the social process, based on their experience of current processes and stakeholder objectives. Excluding further protected group variables narrows the mathematical problem space significantly. The declared overall objective of the ML project for employee turnover prediction is to retain employees through targeted measures (i.e., incentives like wage increases or promotions). This means that the positive class (probability of turnover is high) is more likely to receive benefits and is therefore identified as the favoured outcome. Most importantly, the model should meet the basic requirement recommended in the HR Analytics literature (Speer 2021) and in regulation (Civil Rights Act 1964) not to violate the 4/5 rule for demographic parity, i.e., the adverse impact ratio (AIR) (the difference between selection ratios) needs to be higher than 80%. Once this most important requirement for diversity has been fulfilled, validity should be optimised regarding the prioritisation of diversity and validity.

In iterative predictor removal, the first step is to benchmark against the diversity and validity trade-off with a baseline model that uses all predictors (Rottman et al. 2023). Therefore, no predictors were removed before the first iteration of the automated model development process. The baseline model (Random Forest) with the highest validity out of ten ML algorithm options satisfies the 4/5 rule for age

(AIR_{age} 84.8%) but violates the gender rule (AIR_{gender} 68.7%). Please note that the resulting model is only a starting point and cannot be used for operational decision-making, as it would be a case of algorithmic discrimination. (Goodman and Flaxman 2017). Figure 15 also shows that both post-hoc XAI methods provided in the interface reveal possible unfair biases related to age and gender.



- Y-axis (vertical axis): Mean difference of predicted turnover probability compared to the mean value
- X-axis (horizontal axis): value of the predictor

→ **Bias indication: ML Models predicts higher turnover probability for young men and older women (Δ ~ 10%)**

- Y-axis: Predictor name and value
- X-axis: The turnover probability of the anonymous employee is equal to the sum of all contributions

→ **Bias indication: ML Models finds local differences for the protected feature age and gender**

Figure 15: Post-hoc XAI explanations in the initial iteration. Predictor effects are extracted using SHAP on an individual level and ALE on a group level to reveal unfair bias.

Audit and iterative adjustment in model training (reactive oversight)

Next, the HRM team examines adverse impacts to find possible unfair biases against individuals or groups and to subsequently identify predictors causing unexplainable issues in this regard. For example, the team mentions that part of the lower average turnover probabilities among younger women is caused by successful post-pregnancy reintegration programmes specifically targeted at this cohort. When evaluating the historical background of this programme and its impact, the HRM experts classified these adverse impacts as explainable and in line with human values – and therefore not a critical concern to AF. Instead, eliminating this explainable bias would harm post-pregnancy reintegration programmes and women in general. Furthermore, the predictor effects revealed a significant impact regarding the numbers and ages of children (see in Figure 15, plot B, "Children under 18 true"). These predictive effects appear to differ between women and men, which the HRM experts considered an unjustified bias. Based on this information, the team decided to exclude these predictors and initiate the second iteration of the technical subprocess to develop adapted ML models. These two examples illustrate the possibility of examining nonlinear predictor effects between the subgroups of two protected groups (age and gender) during human-augmented fairness assessments using post-hoc XAI methods. These insights allowed the HRM team to guide the technical team in the development process via feature selection and the separation of explainable adverse impacts and unfair bias (*reactive oversight*). Interestingly, the results of this second iterative automated model development process resulted in a model that satisfies the 4/5 rule for gender (AIR_{gender} 84.4%) but violates the rule for age (AIR_{age} 71.7%). This example

shows the complexity that arises when multiple protected groups need to be studied, and one AIR can rise while the other falls.

When auditing ML models in the second iteration, no evidence of possible unfair bias was found in the interface in terms of post-hoc XAI explanations. Therefore, the teams jointly decided to exclude the predictor with the highest statistical correlation to age (seniority) and start a third iteration of the technical subprocess. Doing so resulted in several models that met the 4/5 rule for both protected groups. Following the above prioritisation, the model with the highest validity was selected, namely a Random Forest model with AIR_{gender} 81.4% and AIR_{age} 81.3%. In operationalising the revised ML model in HR Analytics, the organisation chooses to leave the locus of the final decision in regular audits. HRM decision-makers emphasise the need for the availability of the interface in HRM daily business to check continuously the fairness of the ML model. This is realised on a cohort basis through the historization of forecasts and decisions and a monitor as the fourth component in the interface.

Discussion

With the success of data-driven and evidence-based decision-making in business analytics, there is a risk that HR Analytics will leave behind the ethical responsibilities associated with opening and closing doors for employees if the knowledge and skills of the HR profession are completely replaced by algorithms (Angrave et al. 2016). The background of HRM professionals facilitates a far-sighted and prudent AF assessment, which could also consider the use of qualitative approaches such as employee surveys to broaden the quantitative-only approach limited to historical and possible data-related biases (Rottman et al., 2023). This is of particular concern because evidence from the public sector shows that while business analytics teams exercise discretion and are aware of social values such as non-discrimination, their value sensitivity does not translate into responsible practices. Instead, they use a variety of arguments to distance themselves from or downplay their responsibilities (Fest et al. 2023).

In this study, we propose a socio-technical framework that combines technical approaches for the automated optimisation of AF with a human-machine interface based on XAI. This sociotechnical methodology helps the organisation establish two-way interactions between social and technical experts through defined (sub-)processes and their interconnections (Dolata et al. 2022; Kordzadeh and Ghasemaghaei 2022). The framework takes into account the domain-specific requirements of HR Analytics (no group-specific automated approaches to AF, easy-to-understand visualisation of predictor effects, auditing options in case of high-stakes decision-making, etc.). This emphasises that socio-technical processes can overcome the limitations of automated AF optimisation approaches as well as the mutual limitations of purely social evaluation approaches in complex ML-based HR Analytics scenarios (Downes et al. 2023; Rottman et al. 2023). In the empirical demonstration, the HRM framework enables the AF assessment to capture multiple protected subgroups of employees and nonlinearities. The sociotechnical system allows HRM expertise to inform automated fairness

optimisation approaches by (1) setting acceptable thresholds to limit problem space, (2) separating justified and explainable adverse impacts resulting from unfair biases and (3) indicating the exclusion of certain predictors.

Implications for research

The main purpose of the proposed framework is to guide practitioners towards fair ML-based HR Analytics. However, it is equally important for researchers to achieve nuanced and heterogeneous results, since evidence for bias is also evident the literature. For example, studies in medicine and pharmacy have been found to focus primarily on men ('gender blind' and 'male-biased'), which in turn negatively influences treatment and health outcomes, especially for women (Verdonk et al. 2009). Although, to our knowledge, similar structural biases have not yet been observed in HRM research, the potential for discrimination should be carefully recognised, as ML-based inductive research methods are increasingly being proposed and used (Putka et al. 2018; Cheng and Hackett 2021). For this reason, we echo Mikalef et al. (2022)'s call that AF should be relevant in every study promoting value-creating and transformative impacts through algorithms, even if only on a small scale or purely to discuss its limitations. The proposed framework, based on both IS and social sciences, thus provides an interesting reference for interdisciplinary researchers applying ML-based inductive methods.

Future interdisciplinary research may shed more light on the circumstances under which ML biases can and cannot be mitigated by human augmentation, as human augmentation of predictions sometimes does not seem to mitigate any adverse impacts (Downes et al. 2023). It would therefore be interesting to understand under what conditions, organisational environments and circumstances algorithms and humans succeed in balancing their respective advantages and disadvantages. In addition, the interface is an important mediator for AF assessment, but how interface developers decide what information to display and convey is a decision in itself (Bader and Kaiser 2019). Future research is needed to investigate how the organisation can ensure that developers select all relevant – but not too comprehensive – information. This includes the pressing issue on how to prioritise and select the rapidly evolving options in XAI and automated bias mitigation approaches.

Implications for practice

This study has several implications for practice. First, the proposed framework and empirical demonstration substantiate that adequate explanations, AF and trust are highly interdependent (Zhou et al. 2022). In this context, sociotechnical frameworks for an AF assessment can also be beneficial for organisations seeking to adapt HR Analytics but which are struggling with acceptance issues (Shin and Park 2019). When ethical considerations like AF play a vital role, humans initially have less trust in fully automated systems, and repairing this trust after an error seems to be less effective (Langer et al. 2023) due to lower algorithmic reductionism and lower perceived procedural fairness (Newman et al. 2020). With the sociotechnical approach, trust and acceptance issues might be mitigated as stakeholders

recognise that the arguments of HR professionals are contributing to the output. The interface that XAI visualisations provide is crucial here, as they influence not only behavioural beliefs (e.g., fairness), but also their behavioural intentions (e.g., use and adoption ML) (Haque et al. 2023).

Second, the proposed framework could also benefit practitioners who must adapt to recent and forthcoming legal conditions. For instance, the European AI Act will define new requirements for ML-based decision-making systems (AI HLEG EU 2019). HR Analytics is considered a 'high-risk' application, as exemplified by the recruitment use case. High-risk applications require either enhanced internal or – depending on the specific characteristics of the implementation – external audits. For audits, organisations are asked to (1) document the intended purpose of the AI system in question, (2) provide detailed user instructions, (3) disclose the methods used to develop the system and (4) justify the critical design choices made by the provider (Mökander et al. 2022). We suggest that following the proposed framework will help fulfil the documentation and disclosure requirements when the audit is extended to external stakeholders.

Third, we emphasise the need to invest in interdisciplinary knowledge. For the technical team, in addition to the core knowledge regarding ML, automated bias mitigation approaches (e.g. Mehrabi et al. 2019), post-hoc XAI methods (e.g. Molnar 2022) and foundations of responsible ML development are necessary (Arrieta et al. 2020). For the HRM team, expertise in key HR decision-making processes (selection, development, retention, etc.), understanding of ethical guidelines and stakeholder objectives are core skills (Charlwood and Guenole 2022). In addition, quantitative skills such as basic knowledge of statistical uncertainty, understanding the risk of ML bias, fairness metrics, the ability to correctly interpret XAI visualisations and critical, rational, data-driven thinking are required.

Conclusion

Human-centred HR Analytics that benefit companies and employees alike manifest in AF, but they also enable AF assessment, as automated bias reduction processes are unable to embed human values and social responsibility solely. The sociotechnical framework and empirical demonstration proposed in this paper help understand the complexity of the path to fair and bias-free HR Analytics. It is our hope that this paper will encourage researchers and practitioners in business analytics and related areas to contribute collectively to the still young but incredibly important field of AF assessment.

2.3 Study 3 | Exploring the individual adoption of Human Resource Analytics: Behavioural beliefs and the role of Machine Learning characteristics

Abstract

The technological capabilities of Human Resource Analytics (HRA), enhanced by recent innovations in Machine Learning (ML), offer exciting opportunities. However, organisations often fail to realise these potentials because of a limited understanding of why individuals choose to adopt or disregard respective tools. Prior research on innovation adoption offers preliminary insights but fails to aggregate the determinants of individual adoption into actionable suggestions for decisions in the ML adoption process. Our study applies focused interviews to examine non-ML experts' reasoning for using a specific tool tailored to a public sector organisation, which corresponds to the usual end-user perspective of ML-based HRA adoption. By drawing from the HRA adoption framework, provided by Vargas et al. (2018), we contribute to the literature by identifying relevant beliefs and experiences influencing one's intention to adopt ML-based HRA and by qualitatively linking these beliefs to ML characteristics such as transparency, automation and fairness. For practitioners, we provide actionable guidance emphasising the need to ensure fairness proactively, as interviewees do not consider this aspect when deciding to adopt ML-based HRA.

Introduction

The diffusion of analytics into Human Resources Management (HRM) processes, including talent management, performance evaluation and workforce planning, presents a promising opportunity. Human Resources Analytics (HRA), as it is referred to in this context, is classified as diffusing innovation and describes "a practice enabled by information technology that uses descriptive, visual, and statistical analyses of data related to Human Resources (HR) processes, human capital, organisational performance, and external economic benchmarks to establish business impact and enable data-driven decision-making" (Marler and Boudreau 2017, p. 15). Modern-day technological advances help HR professionals asserting their value when defending against possible displacement by finance or data science departments (Angrave et al. 2016), and they provide support for a wide range of different HR functions (Prikshat et al. 2023b). In recruitment, for example, HRA can be used to streamline processes and achieve greater speed and efficiency (Hunkenschroer and Luetge 2022), whilst in HR development, it helps to identify the link between employee engagement and performance metrics – and thus positively influences them (Davenport et al. 2010). Although traditional HRA encompasses several statistical approaches and methodologies, Machine Learning (ML), such as deep-learning algorithms and Artificial Intelligence, is expected to drive the greatest change in HRM practice. For example, online work platforms such as Uber, Upwork and Deliveroo automate extensive core business processes, ranging from HRM decision-making to execution in the form of selection, compensation and task assignment – all of which is done through ML (Meijerink et al. 2021, p. 2551). In addition, ML-based HRA tools for predicting voluntary employee turnover allow companies to derive retention strategies that not only reduce costly replacements in the short term, but also retain expertise within the organisation, thereby securing a competitive advantage (Chowdhury et al. 2022).

Prior research identifies individual resistance to ML-based HRA which could hinder its success in corporate practice. In contrast to other HRA technologies, sophisticated ML algorithms, for instance, have the disadvantage of being too complex to interpret easily, which subsequently leads to opacity (Kellogg et al. 2020; Langer and König 2023). As more data and multifaceted algorithms become available, a computer learns more complex patterns and "consequently builds its own representation of a classification decision, [which it does] without regard for human comprehension" (Burrell 2016, p. 10). Therefore, algorithms exceed human abilities to understand the system and can generate severe trust issues (Arrieta et al. 2020). Furthermore, prior research observes attempts to manipulate and exploit these advanced ML-based systems, known as "algoactivism" (Kellogg et al. 2020; Meijerink and Bondarouk 2023), and a more general aversion to advanced algorithms, called "algorithm aversion" (Mahmud et al. 2022). Consequently, the successful leverage of the described potential of ML-based HRA critically depends on the ability to convince the individuals of an organisation to use these systems (Di Vaio et al. 2022).

However, as most academic HR literature aims to understand the factors determining the adoption of HRA on an organisation-wide level (e.g., Margherita 2021), and irrespective of the specific tool (e.g., Vargas et al. 2018), there is very little knowledge on the successful individual adoption of HRA – and especially ML-based HRA. Coming from the apparent need for a better understanding of the individual adoption process for ML-based HRA, as well as the ambiguous effect of ML characteristics, we ask the following two research questions:

RQ1: What beliefs and experiences influence the individual's intention to adopt ML-based HRA?

RQ2: How do the characteristics of ML engender these behavioural beliefs?

To answer these research questions, we examine the individual opinions and thoughts of employees of a public sector organisation about a specific ML-based HRA tool for predicting voluntary turnover, the implementation of which the organisation is currently evaluating. Drawing from the focused interviews method provided by Merton and Kendall (1946), we discuss the performance of the predictive HRA tool, as well as several explanatory figures, with employees in interviews and analyse their personal perspectives, experiences and spontaneous reactions to these different approaches. Following Vargas et al. (2018), we then interpret our empirical results with the help of a conceptual framework derived from the *Theory of Planned Behaviour* (TPB) by Ajzen (1991). On the one hand, our results show that the perceived (self-) efficacy of interviewees also highly depends on the design of the HRA tool and the entered dataset, in addition to perceived skills and competencies. On the other hand, the attitude of the interviewed employees is not only formed by their personal enjoyment or concerns in terms of working with the tool, but also by the way in which they perceive it assists them in their daily work. We additionally identify that several ML characteristics (perceived self-learning capabilities, degree of automation, transparency and trialability) influence behavioural beliefs and in turn effect the adoption of the tool in HRM processes.

Our study makes three main contributions to the literature. First, it contributes to the ongoing debate about the relevant factors driving the decision to adopt HRA (Coolen et al. 2023). By examining this decision from an individual instead of an organisation-wide perspective, we provide deeper insights into the different behavioural beliefs determining the decision to adopt ML-based HRA. Based on our findings, we propose several ML-related extensions and adjustments to the more general adoption framework of Vargas et al. (2018). Second, our study contributes to the current literature on ML design approaches and their effect on HRA adoption (Marler and Boudreau 2017, Langer and König 2021, Haque et al. 2023), and third, it contributes to research on ML transparency, suggesting that appropriate visualisation influences end-user adoption Haque et al. 2023. However, in contrast with Haque et al. 2023, our results demonstrate a lack of ethical reflection, as fairness plays no role in individual decisions to adopt ML-based HRA, albeit protected group differences were made apparent in the interviews.

The paper is organised as follows. The second section reviews the literature on the (individual) adoption of HRA, highlights the related limitations and derives the conceptual framework of our study. The third section summarises the research method, empirical environment as well as the research object and data analysis process. The fourth section presents the results. A refined model at the end of this section summarises the factors that influence individual intentions to adopt ML-based HRA and the impact of ML characteristics. Finally, in the fifth section, the results are discussed and propositions made before a conclusion is drawn.

Related research and theoretical framework

In the following section, a conceptual framework for the present study is derived by summarising and discussing the state of knowledge on the (individual) adoption of HRA.

Prior research regarding the adoption of HR Analytics

The factors that drive or hinder the adoption of HRA have been almost exclusively explored from an organisation-wide perspective (e.g., Margherita 2021; Böhmer and Schinnenburg 2023; Coolen et al. 2023). Prior research draws from the TOE framework (e.g., Pumplun et al. 2019; Chatterjee et al. 2021; Neumann et al. 2022), with the underlying idea that the adoption of HRA from an organisation-wide perspective is mainly driven by technological, organisational and environmental contexts. Technological contexts include, for example, the existing IT infrastructure of an organisation (Neumann et al. 2022), while the environmental contexts can be, for example, competitive pressure or customer readiness (Neumann et al. 2022). The organisational context includes cultural aspects (such as the culture of innovation or change management) as well as resources (e.g., budgets or human capital) (Neumann et al. 2022). Prior research concludes that the employees themselves – aligned with their skills and knowledge – play a major role in the adoption of HRA in corporations (Coolen et al. 2023; Di Vaio et al. 2022). Furthermore, work ethics (Basu et al. 2023) or supervisor support (Prikshat et al. 2023a) have been identified as additional major drivers for organisation-wide adoption.

To the best of the authors' knowledge, only Vargas et al. (2018) have examined the individual adoption of HRA and proposed a comprehensive framework in this regard. Drawing from the *Theory of Planned Behaviour* by Ajzen (1991) and the *Innovation Diffusion Theory* posited by Rogers (2003), the authors explain the actual level of adoption of HRA through an individual's perceived self-efficacy, attitude and social influence regarding its use as well as trialability. Self-efficacy represents an individual's beliefs about their abilities to reach a behavioural goal (Bandura 1977), which translates to their evaluation of the technological and quantitative skills they deem necessary to adopt HRA. One's attitude towards a specific behaviour is derived from the expected consequences of this behaviour (Fishbein and Ajzen 2010). As the perceived consequences of using HRA partly depend on an individual's self-efficacy regarding the use HRA, the latter will influence their attitude, among several other beliefs for the given

context. Social influence represents the perceived norms in favour of or against HRA, and trialability encompasses beliefs about the degree to which HRA can be tested before adoption. Vargas et al. (2018) distinguish the three different decision-making steps of knowledge-gathering, persuasion and decision, whereby perceived self-efficacy is formed during the knowledge-gathering step, and attitude, social influence and trialability are derived during the persuasion step. The conducted survey empirically supports the proposed causal relationships as well as the effect of technology self-efficacy.

Limitations of the HR Analytics conceptual framework

While the derived conceptual framework provided by Vargas et al. (2018) extends the fundamental understanding of individual HRA adoption, it does have some limitations. First, it only includes trialability as a potential technological factor to distinguish between different HRA technologies. The proliferation of ML questions the reality, as the framework does not distinguish between the different characteristics of the HRA tool. Furthermore, as HRA includes many different algorithms, systems and methods (Meijerink et al. 2021), and prior research in information systems finds significant effects of an IT systems's design on its subsequent use Haque et al. 2023, there is clearly the need to further characterise and differentiate the proposed model from this perspective. Especially in the context of ML, research has emerged in the HR (Langer and König 2021), management (Glikson and Woolley 2020) and information systems (Arrieta et al. 2020) literature arguing that transparency must be another fundamental determinant of individual ML adoption. In contrast to traditional statistical methods in HRA, transparency is not always present in ML, because (a) predictors are not understandable, (b) relationships between predictors and predictions are hidden and (c) no explanation for a specific prediction is given (Arrieta et al. 2020; Burrell 2016; Langer and König 2021). This is problematic, because a prediction without clear explanations, or at least justification for the rationale behind the prediction, can lead to trust issues (Glikson and Woolley 2020; Langer and König 2021). Park et al. (2021), for instance, illustrate that only with sufficient transparency can various user burdens (emotional, mental, biases, etc.) be overcome during ML adoption. Transparency is also closely related to another fundamental determinant of HRA adoption, namely fairness (e.g., no discrimination against minorities), which can only be tested when professionals use their expertise and experience to determine the level of fairness of individual ML predictions through intuitive thinking (Chowdhury et al. 2022, p. 18). To achieve sufficient ML transparency, Explainable Artificial Intelligence (XAI) offers a rapidly evolving interdisciplinary research area with multiple technical solutions (Arrieta et al. 2020). Finally, the ability to automate decisions fully is an ML characteristic that represents a major shift for traditional HRA technologies (Meijerink et al. 2021, pp. 2545–2546). When algorithms are used for automated scenarios, they must also be accountable for the decisions they make (Busuioc 2021, p. 826). In summary, and in line with Lee and Cha (2023), we suggest that transparency, fairness and accountability (in terms of automated use) determine the adoption of ML-based HRA. Besides technical ML characteristics, decisions made during roll-out also affect the adoption of ML. In this regard, some studies have found

that the ability to try (trialability) has a positive impact on adoption (Omrani et al. 2022). However, research is still inconclusive in terms of exactly how these ML characteristics influence an individual's adoption of HRA.

Second, the notion of self-efficacy and attitude in the proposed framework of Vargas et al. (2018) is relatively narrow, and it might exclude potentially relevant beliefs. Compared to assumed self-efficacy, the perceived behavioural control (PBC) factor from the original TPB is a wider concept that includes beliefs about factors beyond one's individual control (Ajzen 1991). It can be defined as the perceived ease or difficulty of performing a behaviour (Ajzen 2002). For instance, the individual adoption of an HRA tool likely depends on the tool's suitability for a task and not only on one's perceived skills to use it. Furthermore, Vargas et al. (2018) examine self-efficacy regarding technology and mathematics in general, which are sufficient to estimate the average intention to adopt HRA but fall short when comparing the adoption of different HRA systems. However, the TPB is built upon the principle of compatibility, which states that the underlying factors must always refer to the underlying behaviour (Fishbein and Ajzen 2010). For the given context, one would therefore expect a notion of self-efficacy that is more directly connected to the individual adoption of a specific HRA tool or system. Furthermore, the attitude of a survey participant is derived from four beliefs solely centred around the personal enjoyment of using HRA (Vargas et al. 2018). This notion contrasts with the *Technology Acceptance Model* that connects the attitude towards a technology to the beliefs about the perceived usefulness and perceived ease of use of a technology (Davis 1989) and the *Unified Theory of Acceptance and Use of Technology* that connects the respective attitude to a performance and effort expectancy (Ajzen 2002; Venkatesh et al. 2003).

Underlying conceptual framework theory

Due to the limitations of the conceptual framework of Vargas et al. (2018) described herein, we aim to scrutinise the framework and extend it to ML-based HRA tools. Our further analyses are based on the assumptions described above, which are founded on the current state of knowledge. We also distinguish between the process steps of knowledge, persuasion and decision (Rogers 2003) in an ML-based HRA tool's adoption process. In the knowledge step, personal beliefs are evaluated regarding the ability to utilise an ML-based HRA tool for a given task and form an expectation about the PBC. In the persuasion phase, personal beliefs are evaluated regarding the consequences of using the provided HRA tool and form a tool and task-specific attitude. In addition, personal beliefs are evaluated regarding the opinions of others regarding the use of the provided HRA tool and form an expectation about the relevant social norm (corporate or national culture). In the decision step, personal beliefs are evaluated regarding the PBC, attitude as well as perceived norm and help decide whether to adopt the provided HRA tool. Furthermore, we expect PBC and attitudes to be influenced by the technical characteristics of the

provided ML-based HRA tool, in which case we distinguish between the known characteristics of trialability, transparency, degree of automation and fairness and potential unknown characteristics.

Research approach

Method

To fill the derived conceptual framework for individual adoption with salient beliefs, it is necessary to dive deep into the line of reasoning employed by end-users. We aim to explore these beliefs by applying the "focused ethnographic interview" methodology proposed by Merton and Kendall (1946). We opted for a qualitative research approach because it can provide new insights into individual adoption in an explanatory manner. In addition, the open-ended nature of the interview questions allows for the collection of a wide range of information, including personal perspectives and experiences. The interview procedure was semi-structured around several pieces of information and nudges, used as potential triggers for spontaneous reactions. During the interview, detailed discussions were held on hypothetical but realistic implementation scenarios for the specific HRA tool. Particular attention was paid to employees' understanding of the presented tool, their ideas about its future use in HRM processes and their perceptions of the risks and benefits of using it in various HR applications throughout the organisation. In addition, the interviews provided information about the overall intentions of the interviewees as well as any changes in their intention to adopt the HRA tool when providing various information and explanations. This approach follows the interpretive tradition of explorative methods, in that it seeks a deep understanding of human experience rather than rigid explanations of cause and effect – as in positivist epistemology (Einola and Khoreva 2023, p. 121).

Empirical environment

This study examines a German federal agency from the social insurance industry with about 20,000 employees in the period between 2022 and 2023. The in-depth public sector study approach provides a context in which high legal requirements for the individual adoption of HRA can be investigated and commercial secrecy is not a concern (Desouza et al. 2020, p. 206; Busuioc 2021, p. 826). While the organisation frequently uses descriptive analytics based on advanced dashboarding tools, as well as sporadic diagnostic regression-based analytics, this project is the first to incorporate complex ML models to implement predictive analytics use cases within HR.

Our main objective in selecting the interview population was to obtain a diverse sample of HRA users in terms of personal characteristics (age and gender), seniority and statistical background in order to represent the diverse workforce of the organisation as well as the different usage objectives in the different personas. Table 6 provides an overview of the 12 interviewees. Team leaders supervising one to 21 employees, and heads of departments with 21 to 50 employees, from HR and operational departments, are the main users of the HRA tool. Half of the employees interviewed would work with

the HRA tool in the near future, and half of those interviewed were potential recipients for further applications. Each interview lasted between 58 and 98 minutes.

	Organisational section	Department	Position	Sex	Seniority(y)
I1	Corporate development	Employer branding & image	Team lead	w	20 to 25
I2	Internal corporate consultancy	Management of future vacancies	Team lead	m	15 to 20
I3	Insurance claim processing	Operational workforce management	Department administration	w	30+
I4	Insurance claim processing	Operational workforce management	Head of department	m	30+
I5	Human Resources	Organisation design	Organisational consulting	m	30+
I6	Human Resources	Organisation design	Team lead	m	0 to 5
I7	Human Resources	Organisation design	Department administration	w	25 to 30
I8	Human Resources	Personell planning & controlling	Associate	m	10 to 15
I9	Human Resources	Personell planning & controlling	Project lead	m	15 to 20
I10	Human Resources	Recruiting, development & diversity	Team lead	w	30+
I11	Human Resources	Strategic workforce planning	Senior data analyst	w	0 to 5
I12	Human Resources	Strategic workforce planning	Analyst	m	20 to 25

Table 6: Interview population of future HR Analytics users

Research object

The specific ML-based HRA tool investigated herein predicts individual voluntary turnover (excluding age-related reasons and termination on the part of the employer) probabilities within the next 6 months for each employee, using the random forest algorithm (Breiman 2001). The tool is trained on a fully anonymised dataset with monthly data over a three-year time horizon and includes 30 predictors originating from the same federal agency in which the interviews were conducted. Work-related predictors include commuting distance, sick days, salary, salary increases in recent years, seniority and others. Demographic data such as gender, age, number of children and education level are also included. The ML predictions are evaluated in an out-of-sample test dataset. Instead of treating the ML model as a black box, post-hoc XAI explanations at the local (employee-specific) and global (organisation-wide) levels are used to extract the effects of the predictors. The confusion matrix used to assess predictive accuracy, as well as some visualisations of the XAI results at the local and global level, were used as nudges during the interviews (see Figure 16). The visualisation of organisation-wide explanations describes how a single predictor influences the employee turnover prediction (strength, positive/negative contribution) on average, considering all employees in that local interval (Apley and Zhu 2020). Additionally, the visualisation of employee-specific explanations breaks down the probability of voluntary turnover for each employee and quantifies it in terms of increasing or decreasing effects. The mean value represents the average employee turnover risk of all employees predicted in the model.

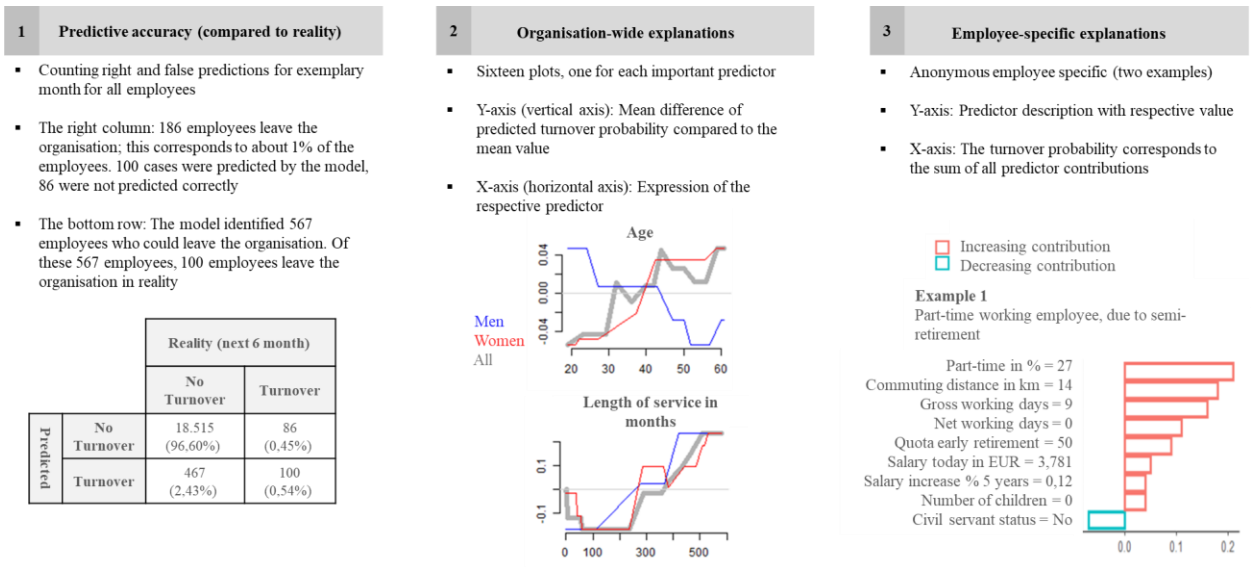


Figure 16: Information provided as nudges during the interviews: Predictive accuracy report, predictor effect explanations on the organisation-wide and employee-specific level

Data analysis

Given the inductive nature of the study, we coded the transcripts manually, following the methodology proposed by Gioia et al. (2013), which has demonstrated its validity in numerous renowned publications over the last decade (e.g., Friedman and Ormiston 2022; Schuessler et al. 2023; Mula et al. 2024). All interview transcripts were coded independently by the first and second authors using MAXQDA[13] software. A total of 392 codes resulted. In the initial coding phase, we strictly adhered to the terms, phrases and descriptions of the interviewees, so that many first-order categories emerged. After eight interviews, both authors re-checked their coding to improve the reliability of the process and increase its rigour and authenticity. After coding all transcripts, first-order categories were compared against each other. Disagreements regarding interpretation, and thus coding, if any, were resolved through discussion. In the next step, for the second coding (axial coding), the first and second authors looked for similarities between and differences among the many first-order categories, in order to summarise and condense them. To this end, we went through each interview transcript as well as the first-order categories again. Subsequently, we discussed each passage and then reconciled different interpretations and conclusions to generate suitable second-order categories (Gioia et al. 2013).

Based on the TPB (Ajzen 1991), each of the second-order categories was independently assigned to PBC, attitude or norm (aggregated dimensions) by the first and second author (as determinants of an individual intention to adopt HRA) and then discussed. Subsequently, all second-order categories were critically reflected in correspondence with the framework provided by Vargas et al. (2018). The third author, who participated directly in project meetings and reviewed relevant project documents, critically

[13] https://www.maxqda.com/

reflected on the results in the final analysis step. Additionally, the coding of the entire interview material was repeated to verify validity. The re-coding of the first author, 11 months after the first coding, resulted in an overlap of 90.4% (intra-coder reliability). Coding by a person not previously involved in the research process resulted in an appropriate accordance of 79.0% (inter-coder reliability) (Miles and Huberman 1994). The interview coding process is summarised in Figure 17.

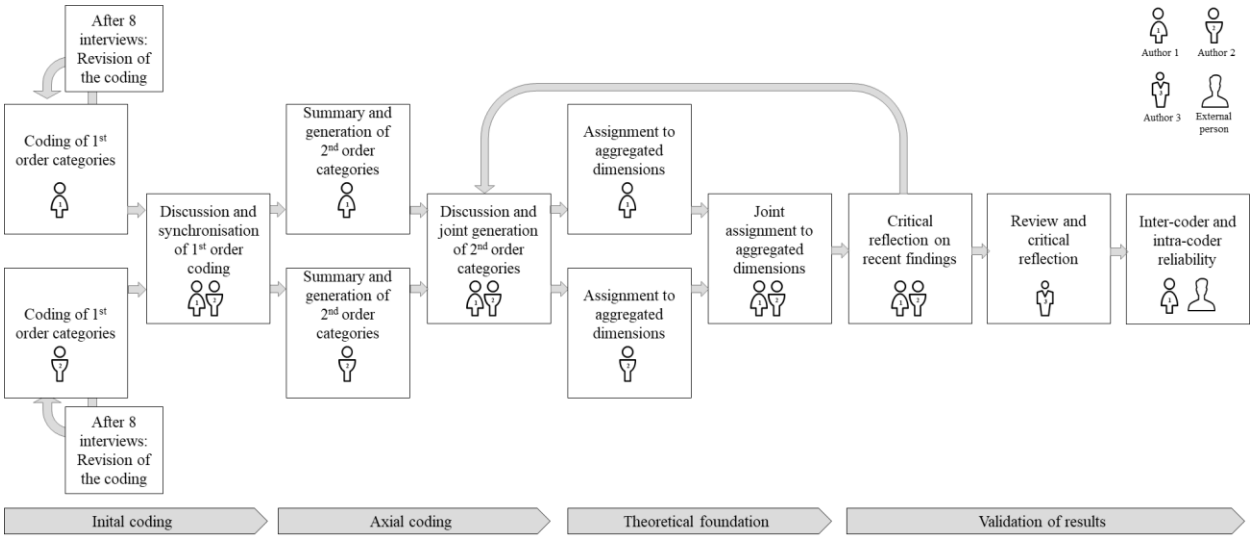


Figure 17: Interview coding process, including critical reflection steps to ensure reliability

The second-order categories and the aggregated dimensions formed the basis for the framework developed for the present study, and the first- and second-order categories, as well as the aggregate dimensions, became the basis for building the data structure (see Figure 18).

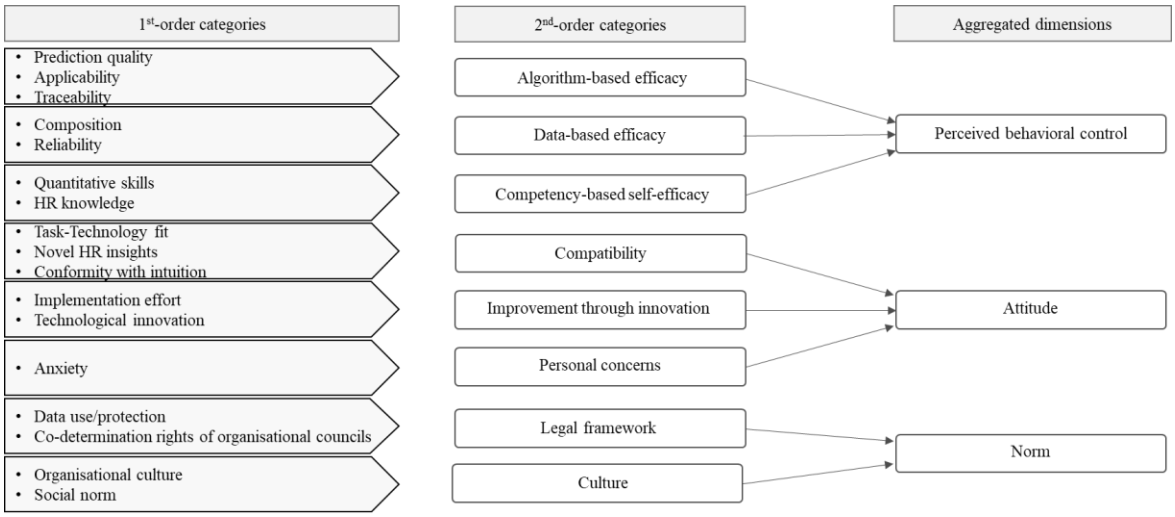


Figure 18: Data structure (first-order categories summarised by key topic)

Results

Factors influencing the individual intention to adopt ML-based HR Analytics

The results regarding the beliefs and experiences that influence individuals' intention to adopt ML-based HRA (RQ1) are presented below. First, the findings on PBC are illustrated (see Table 7), following which, after describing attitudes towards the adoption of the tool (see Table 8), the perceived norms (see Table 9) of the interviewees are presented.

PBC

Algorithm-based efficacy describes all the capabilities and characteristics that interviewees attribute to and expect from an ML-based HRA tool, which allows them to utilise it in their daily work. Prediction quality captures the prediction accuracy of the tool, whereby the interviewees seem to require a sufficient level of prediction accuracy to view it as applicable. For our model, prediction accuracy is rated differently by I1 and I11. Another key factor for algorithm-based efficacy is traceability. The lines of reasoning by I10 reveal that the HRA tool's feasibility stems from explanations of the ML model and how this helps to optimise the organisational retention of top performers at the individual level. I12 points towards the XAI visualisations provided as an important distinction compared to popular Generative Artificial Intelligence models such as ChatGPT. While scepticism towards these technologies is generally high, trust-building and achieving actionable insights can partially be attributed to traceability. The third driving factor for algorithm-based efficacy is applicability in practice. Among other things, the interviewees consider the extent to which the tool is mature and ready for use, its susceptibility to errors and the basic functions (e.g., the selection of different prediction periods) that it offers.

In addition, data-based efficacy plays an important role when forming the PBC. In our case, the opinions of the interviewees regarding the composition of the dataset differed widely. Some interviewees, like I2, identified missing and crucial predictors of voluntary turnover from their point of view, which had a critical impact on their evaluation of the tool. Others, like I9, were very satisfied with the included predictors. If they identified turnover predictors in the data, which they might consider important, data-based efficiency was considered high. Similarly, if the contribution and importance of the predictors in the XAI visualisations are as expected, then data-based efficiency seems to increase. Additionally, we find that dataset reliability plays an important role in the intention to adopt the HRA tool. For instance, I10 directly attributed the reliability of the department responsible for managing the database to the tool. Others, like I11 questioned the timeliness and quality of the data, its realism or rapid changes in the included turnover predictors.

In our study, competence-based self-efficacy reflects the beliefs an individual holds regarding their ability to use the ML-based HRA tool successfully in their daily work. Some interviewees, like I6,

quickly understood the nudges shown, were interested in them and interpreted the information in detail. In their daily work, these people mostly take on analytical tasks and often have a background in statistics, which indicates that they have more pronounced quantitative skills. Others, like I8, were overwhelmed with the interpretation and had no deep interest in the information provided to them. In addition, the interviewees considered it necessary to have a certain level of HR knowledge, to be able to apply the results of the tool in practice and to derive potential application scenarios. The findings on our interviewees' PBC-related adoption of ML-based HRA tools, which are presented in Table 7, match interview insights into the dimensions related to the Theory of Planned Behaviour.

Example Insights	1st-order Categories	2nd-order Categories	Aggegate Dimension
"*... actually, quite impressive prediction quality.*" (I1, 34)	Prediction quality	Algorithm-based efficacy	PBC
"*... if you look at the absolute numbers, 467 and 100, prediction accuracy doesn't look so great.*" (I11, 102)			
"*I understood that the higher the absenteeism due to illness, the higher the probability that these people will leave the organisation, which would allow me as a personnel manager to conclude: How high is my sickness rate? [...] Unfortunately, my department has a very high sickness rate, and it would be exciting for me to see whether this has led to increased turnover – that the sick days were perhaps even the criterion.*" (I10, 104)	Traceability		
"*I do not know how to calculate the predictions. I do not know how the database must be prepared, how the model must be fed and so on. But if I look specifically at the XAI visualisations, I can already work with that. […] You first must deal with it [like a new software program] to be able to use it. [...] A little bit of scepticism is quite healthy, but ChatGPT has now increased our trust somewhat.*" (I12, 171)			
"*I would need 5 or 10 years and not the individual level of an employee, but I would have to look at the entire department, and I would have to look at certain levels, e.g., professional groups.*" (I11, 162)	Applicability		
"*All these flexible working time models with remote working etc. are not integrated into the model. After all, these affect 50% of our employees*." (I2, 108)	Composition of the dataset	Data-based efficacy	
"*These are very important predictors that I could then use for the future. So that would be very helpful, very helpful.*" (Interview 9, 92)			
"*So you certainly calculated this from the data collected by Mr. […], I assume? From there, the data basis is safe for me – and from there I also trust in the numbers.*" (I10, 242)	Reliability		
"*So, for example, the economic situation, the issue of security, the issue of a personal family situation. The factors change. I hire someone who does not have any children, and then I know, well, maybe in 5 years there will be children. That means I cannot exert any influence. Likewise, what about changing health situations?*" (I11, 128)			
"*I understood the figures shown.*" (I6, 85)	Quantitative skills	Competency-based self-efficacy	
"*But at the moment, it slays me and everything – honestly.*" (I8, 249)			
"*I cannot say anything about that [how to adopt the tool specifically]. My colleagues in the HR department are more closely involved in this issue. I cannot assess the potential.*" (I12, 167)	HR knowledge		

Table 7: Interviewees' PBC related adoption of ML-based HRA tools

Attitude

The interviewees' attitude towards the adoption of HRA was influenced, among other things, by its perceived compatibility. The assessment of the task technology fit was very different (see I3 and I7), with the added value and the concrete integrability of the ML-based HRA tool in everyday work being questioned and analysed. The interviewees' attitudes also seemed to be influenced by whether the results of the HRA tool revealed novel HR insights. Like I9, almost all interviewees stated that the tool could

help identify at least some novel factors for employee voluntary turnover – and thus provide a basis for the development of personnel measures. It is notable that the consistency of the tools' predictions with personal intuition is an important determinant of attitude. Provided that the results matched the intuition, this manifested in an improved attitude, and vice versa.

In addition, considerations related to the improvement achieved by the tool seemed to have an impact on attitude. In our study, the extent to which implementation effort and technological innovation were perceived as impactful by the interviewees was important in this context (see Table 7). Some, like I4, felt that the innovative nature of the HRA tool enables new approaches to old challenges (such as demographic challenges) and improves previous processes (e.g., the quality of workforce planning). Essentially, innovation brings new perspectives and approaches. Others, like I11, critically questioned implementation efforts in terms of a cost-benefit trade-off.

A few interviewees questioned the personal consequences of adopting the HRA tool and evaluated them accordingly. I9 and I11 were particularly afraid that superiors use the HRA tool inappropriately (e.g., findings led to monitoring by the superior or mobbing), that false predictions led to negative effects (e.g., in the allocation of tasks) or that misunderstandings occurred. Findings relating to the interviewees' attitude to the adoption of ML-based HRA tools are presented in Table 8.

Example Insights	1st-order Categories	2nd-order Categories	Aggegate Dimension
"*I just wanted to add, because I think that our organisation is so big, that you do not have to look at individual employees. I cannot use that method at all.*" (I3, 109)	Task technology fit	Compatibility	Attitude
"*Through the tool, managers are encouraged to be active in their role.*" (I7, 362)			
"*Then I can look at this and analyse the most important factors that influence why this employee is leaving and use this to initiate optimisation.*" (I9, 120)	Novel HR insights		
"*Of course, this creates trust when you see that even without algorithms.*" (I3, 64)	Consistency with intuition		
"*No, there must be a mistake. It says that the probability of turnover is higher for civil servants.*" (I5, 115)			
"*I see this as a great support, and it goes much further than what we could do in the past. […] You can draw insights from the data that give the organisation a positive kick in any case.*" (I4, 283)	Technological Innovation	Improvement through innovation	
"*I think it is great that such an approach has been found at all*." (I6, 160)			
"*It is always nice to try something out, but of course, the question is then always cost and benefit. Does it bring us anything?*" (I11, 162)	Implementation Effort		
"*I am afraid of the surveillance now that the supervisor monitors me like this: do I go or not?*" (I11, 206)	Anxiety	Personal concerns	
"*On the negative side, my responsible tasks could be taken away from me, because there would be a risk that I would leave the organisation.*" (I9, 212)			

Table 8: Interviewees' attitude to the adoption of ML-based HRA tools

Norm

The adoption of (ML-based) HRA tools is also limited by the 'legal framework'. The interviewees stated that possible applications of the tool were severely limited or not possible due to legal conditions and the strict interpretation of data protection regulations in the public sector. Interestingly, some interviewees mentioned experiences with – from their point of view – overly strict data protection rules for historical organisational initiatives. For example, I8 stated that sensitive personal data should also be included in the tool and used for individual decision-making. Organisational councils have far-reaching co-determination rights that go beyond the law and enable employee representatives to object to various decisions affecting the entire organisation, thus obstructing adaptation. The interviewees perceived that initiatives based on employee data were – in principle – prevented. Among other things, organisational councils receive evaluations of all HR reports requested in the IT system, and they strictly ensure that each employee processes only the amount of information needed to complete tasks.

Culture is an important factor in the adoption of HRA (Vargas et al. 2018), where we distinguish between the social norm and organisational culture. During our interviews, some interviewees critically evaluated their social norm's compliance with the adoption and use of the HRA tool. For example, in terms of the individual employee's privacy, they questioned whether the analysis of personal data was acceptable from their point of view, or with whom the responsibility for ensuring appropriate use lay. Others, like I11, saw no threat to (personal) privacy. From the interviewees' responses, we were also able to find indications of the anchored organisational culture, which in our case tends to have a hindering effect on the HRA tool. The interviewees stated that there were many people with reservations and sceptics who viewed changes to previous processes or systems as negative; in addition, decision-making processes within the authority were often perceived as not rational and were very time-consuming. Moreover, the adoption of the HRA tool was a complicated undertaking because employees had difficulty dealing with predictions and uncertain expectations. Interview insights on subjective norms regarding adoption of ML-based HRA tools are presented in Table 9.

Example Insights	1st-order Categories	2nd-order Categories	Aggegate Dimension
"*Data protection is a very important topic in our organisation. [...] Because we run analyses here that can be evaluated on a personal basis, and conclusions can be drawn about a person* " (I6, 180) "*This is very sensitive data with which the tool works – highly explosive in terms of data protection. Therefore, it cannot be implemented in this form.*" (I7, 326) "*And when it comes to data protection […] I think [it] tends to protect those who have something to hide rather than benefit others. […] Instead of excluding variables, you might have to take other variables in addition*." (I8, 351-359)	Data use/protection	Legal Framework	Norm
"*The problem is: The implementation of tools like this in-house must be approved by the organisational councils. From my work as an organisational consultant, I also know that software like this is not simply approved*" (I5, 169)	Co-determination rights of organisational councils		
"*That would just be too much intrusion into my personal life for my supervisor to have that information to hand.*" (I9, 258) "*My boss has all my data at his disposal. He knows how many children I have, and he also knows where I live. He also knows when I'm sick and how much I earn.*" (I11, 214)	Social norm	Culture	
"*We have many doubters – there is not only the political thing in the house.*" (I7, 366) "*The management always tries to be supportive, of course, but decisions are not made as quickly as in the private sector.*" (I9, 240)	Organisational culture		

Table 9: Interviewees' subjective norms regarding the adoption of ML-based HRA tools

Impact of ML characteristics

The results for how the characteristics of ML affect behavioural beliefs (RQ2) are presented below. Our findings are successively illustrated in terms of trialability, transparency, automation, self-learning capabilities and fairness, as well as their effect on beliefs and experiences.

Trialability has an impact on attitude

Overall, we observe a positive effect of trialability on the intention to adopt the provided HRA tool. Similar to the findings of Vargas et al. (2018), our interviewees believe that it is important to try out the ML model before it is implemented in the organisation, in order to gain experience of using it. They argue that a high level of trial and error makes it easier to assess the accompanying consequences of actually applying the ML model, which in turn could lead to an increase or a decrease in one's attitude regarding the tool. On the one hand, trialability helps to assess whether the ML model provides a presumed improvement through innovation (technological innovation):

> "*I like to try something like this out in practice […]. ML does not really help here yet. I think we always have to make our own experiences with applications. […] They have to prove themselves in practice somewhere. And if they do not, then I have to analyse that. Where is the problem, or where does it not bring the benefit that I had hoped for? And, if necessary, I have to adapt it.*" (I4, 283)

On the other hand, trialability helps to mitigate any potential personal concerns of employees:

> "*For matters that are more critical, it is wise to first try things out, test them, see where adjustments can be made, involve the people and initially test it in a small area to then see how it is received [...] But having these sceptics around all the time makes everything a bit more difficult.*" (I11, 222)

ML transparency has an impact on both attitude and PBC

We observe positive and negative effects of transparency on the intention to adopt the provided tool. At the beginning of the interviews, we asked the interviewees about their intention to adopt the employee turnover predictions in their daily work. Interestingly, most initially saw little to no application in the tool's predictions when it came to pure predictive accuracy without understanding the effects of the predictors:

> "*Unfortunately, I am not able to determine the value added because I have not performed any [proving] calculations [...]. Therefore, I could honestly plan better for the future based on historical data.*" (I9, 98)

However, the more transparency provided by the presentation of multiple predictor effects, the more diverse and extensive the applications identified by the interviewees (in their areas of responsibility), improving their attitude via the perception of compatibility of the tool and especially the personal task-technology fit. Besides reflecting on how the predictions could be used (e.g., for workforce planning and identifying future staff shortages), the interviewees also recognised that the tool provides explanations for turnover. Thus, it offers opportunities to either mitigate turnover at an individual level or derive strategic and organisation-wide initiatives that address employee wellbeing (e.g., increasing remote working opportunities) and employer attractiveness (e.g., increasing childcare offerings). The discussions in all interviewees about possible applications of the tool in other HRM processes, made possible by transparency, indicate a higher algorithm-based efficacy.

In addition, providing more transparency can have a positive or a negative effect on one's attitude when the derived predictor effects contradict personal intuition (compatibility). On the one hand, our interviewees found contradicting evidence useful in questioning their personal intuition:

> "*But it definitely brings insights that straighten out the picture and probably bring it closer to reality. Yes, I would use it if I had to decide for my hotdog stand.*" (I1, 135)

On the other hand, a few interviewees questioned the functionality of the provided tool when identifying evidence that contradicted their own intuition (conformity with intuition, Table 8). Furthermore, our results suggest that these interviewees demanded a high degree of traceability to help them understand the underlying calculations of the ML model (algorithm-based efficacy):

> "*I [...] want to understand what is happening behind the system, [...] In Excel, you can see how the calculation is done and what the result will be. With machine learning, you probably won't be able to see it that way. The machine learns based on the data and then outputs something. So, I always need a certain level of traceability for each step.*" (I9, 294)

To summarise, we find that transparency influences attitudes via the perception of compatibility in two ways (personal task-technology fit and conformity with intuition), as well as PBC via algorithm-based efficacy, also in two ways (applicability and traceability).

Degree of automation through ML decision-making influences attitude and PBC

We mostly observe a negative effect of the degree of automation on the intention to adopt the provided ML model. All interviewees agreed that decisions should only be augmented with the help of the tool and that a fully automated decision-making process should not be implemented. Several reasons regarding attitude, especially personal concerns, were given for this, such as the fact that the interpersonal component must not be lost, especially when decisions are made on an individual basis:

> "*At the top level, they want numbers, and there's also the risk that when they see those numbers, they do not want to deviate from them […]. However, the human factor, and the perspective and the focus on the individual employees, is simply lost as a result. The decision-makers who normally have management responsibility, who actually manage people, have to look at the results.*" (I1, 52)

I4 pointed out that automation is only useful if the model does not make a single mistake. This in turn is reflected in the expected accuracy of the tool (algorithm-based efficacy):

> "*[For automation], the probability of correct predictions is not yet high enough, not until the hundred per cent mark is reached. Until then, decisions are up to personnel analysis by management – instead of letting the machines think completely.*" (I4, 275)

The interviewees believed that the responsibility and rational towards decisions lies with humans (an ML-augmented decision process) and questioned whether the provided ML model is suitable for drawing the right conclusions and deriving appropriate actions from a prediction. This translates to a low perceived applicability (algorithm-based efficacy):

> "*If you were to go only by the machine: a woman has a salary of 3,000. We will just raise it to 4,000 – but a man does not get that raise. [...] I would see it critically in the first instance. In any case, it does not replace the interpersonal connection. Well, I do not work together with the machine, especially not in a subordinate relationship. Ultimately, such a decision must be made by a manager.*" (I7, 402)

The self-learning capabilities of ML affect PBC

Unexpectedly, we identified the perceived learning capabilities of the provided ML model as a further relevant ML characteristic influencing the algorithm-based efficacy (PBC). A few interviewees associated continuous learning with ML and expected continuously increasing accuracy due to future learning iterations with more data or feedback loops:

> "*[With] ML and Artificial Intelligence work – as far as I have now generally heard – the more you feed, let's say, the machine with information, the better it becomes. And that's exactly the direction it should go if you use it more often and feed it with more and more data. It will get better and better, and that will also reduce the error rate, in my opinion.*" (I9, 182)

Interestingly, some of the interviewees translated the automated self-learning characteristics they were aware of from a reinforcement learning ML model in another context to this specific ML model, without knowing whether these feedback loops were actually implemented:

> "*[ML]… is a self-learning system, and the more often I run it, the better my predictions become. In this respect, if I have understood correctly, we are still at the start. And the more data is fed in and compared with real things, the more accurate the predictions will be – at least that's what I would expect.*" (I2, 85)

(Un-)Fairness does not affect the intention to adopt

As noted in the literature review, increased ML transparency can also affect perceptions of ML fairness. In the interviews, we specifically asked about the fairness perception in hypothetical scenarios (e.g., What requirements do you have for the model in terms of fairness or equal opportunity? Would you remove certain data from the dataset, for example, for reasons of fairness or equal opportunity? From the position of your supervisor making decisions about you, do you have any reservations or concerns about using the model?). Interestingly, none of the interviewees mentioned significant caveats regarding fairness aspects. For example, excluding protected group variables from the dataset was not suggested by any interviewee. We were able to ascertain this even after providing a list of predictors used in the ML model as well as XAI visualisations that (1) listed protected group variables such as gender or age and other data that require extensive protection, such as health-related information, (2) clearly documented differences between protected group variables and their impact and (3) were considered the basis for local (employee-specific) or global (organisation-wide) decision-making – and thus varying degrees of impact on individual employees. Several arguments were made by the interviewees as to why the HRA tool in its current state is fair and no adjustments are needed. First, I9 referred to the objectivity of the data and the responsibility for any consequences:

> "*You cannot influence the fact that our organisation is over 70% women, so you cannot do anything by saying that we now have to hire only men. That would be discriminatory. That is*

why I do not think it is so bad. Those are the facts, that is the database, you cannot change that. Or that older workers are less likely to quit. As a human being, as a decision-maker, you also have the information, and you have to interpret it accordingly." (I9, 280)

Second, I1 made a similar argument, pointing out the possible lower prediction accuracy of the tool when predictors are eliminated. The interviewee argued that differences are not unfair if they are based on differences between protected groups that can be explained by real facts, which he exemplified in terms of intergenerational differences:

"*My father was in the same company for 30 years, my mother worked in the same company for over 40 years. So that is the thing, for me, that would be unthinkable. I would not exclude predictors like age or gender […]. I can understand that you want to leave such factors out so as not to discriminate against anyone […], but it would also just be out of touch with reality. […]. I would claim that there are significant differences, and that tells me that it must not be left out at all.*" (I1, 151)

Overall, we found no evidence that the interviewees had changed their intention to adopt the HRA tool due to the different treatment of women or men, or younger and older employees. This can be explained by the above statement that potentially unfair discriminatory decisions should be corrected by human judgment in an augmented (non-automated) decision-making process.

Refined model: Individual intention to adopt HR Analytics

The results of our study are summarised in a qualitative model, as illustrated in Figure 19. Vargas et al.'s (2018) framework forms the basic structure on which the various factors influencing PBC, attitude and norms are concretised. As a further addition, the influences of ML characteristics emerge. Please note that ML characteristics have effects on zero (fairness), one (self-learning capabilities) or two (transparency, automation and trialability) constructs of behavioural beliefs.

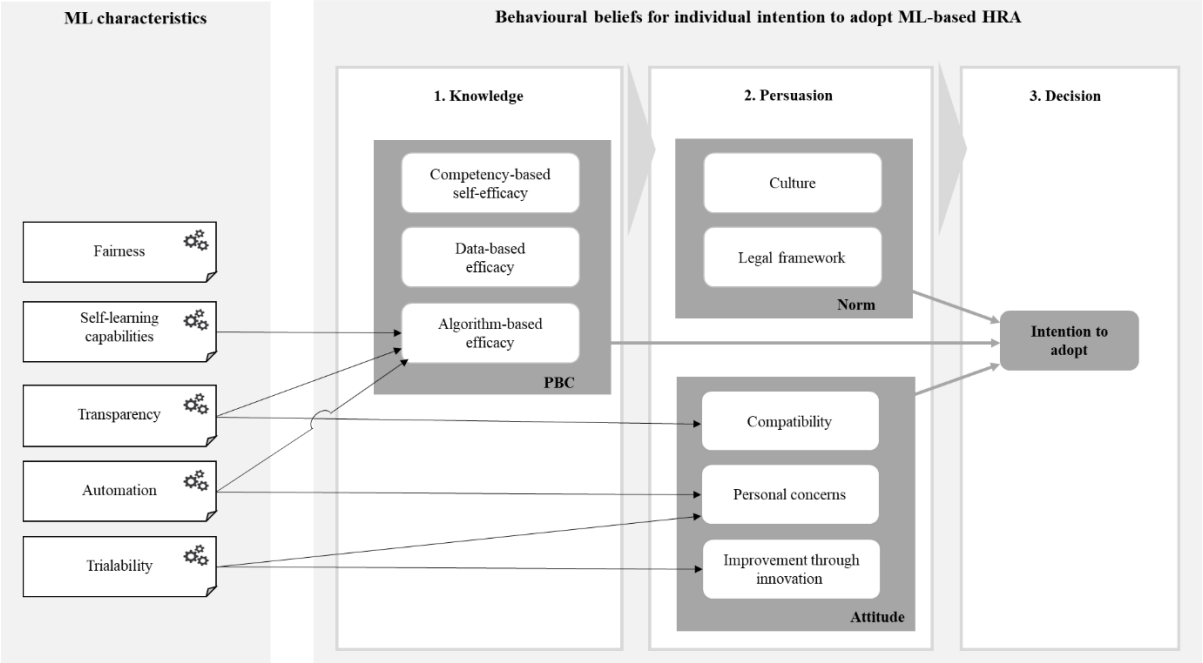


Figure 19: Proposed qualitative model for individual intention to adopt HRA based on ML characteristics.

Discussion and implications

Notwithstanding the asserted importance of HRA, research examining its impact on organisational performance remains underdeveloped (Marler and Boudreau 2017, p. 15). While recent studies find evidence for the effects of HRA and organisational performance, this link is mediated by an organisational shift to more evidence-based management practice. Moreover, it is argued that HRA can only provide a benefit for an organisation when predictions and estimations are incorporated into neutral and evidence-driven decision-making (McCartney and Fu 2022, p. 38). Evidently, a respective shift requires a congruent employee mindset and therefore a strong intention to incorporate HRA into their daily work. Our study contributes to this ongoing debate by extending and contextualising current knowledge on the adoption of ML-based HRA for a specific use case.

PBC, attitude and norms influence an individual's intention to adopt HRA

Regarding salient beliefs and experiences, Vargas et al. (2018) find five factors determining the individual adoption of HRA: technology self-efficacy (PBC), quantitative self-efficacy (PBC), attitude towards using HRA, social influence and tool trialability. Vargas (2016) investigates a larger number of possible factors influencing the user level of adoption, whereby general self-efficacy and data availability, which were queried in the survey with narrowly specified items, showed no significant influence. Therefore, we investigate the personal perspectives, experiences and spontaneous reactions of knowledgeable employees in a specific use case with field data. This particularly provides a lens for a deeper understanding of the individual adoption of ML-based HRA. For example, we find evidence that interviewees with higher quantitative skills, as well as overall competency-based self-efficacy (e.g.,

HR knowledge), were more open to incorporating the displayed HRA tool in their work. In addition to technological aspects, the input data on which the ML-based HRA tool is trained was highly important to the interviewees. We therefore suggest that data-based efficacy, as an aggregation of the composition and reliability of the dataset, is another important dimension of a potential user's efficacy. The findings published by Omrani et al. (2022) suggest similar relationships in terms of data, albeit their study argues that concerns about discrimination in the use of Artificial Intelligence reduce trust. In addition, we find evidence that attitudes toward the adoption of HRA are made up of a wide variety of aspects. These can be divided into three categories, namely the assessment of the tool's compatibility in daily work, an assessment of the potential for improving the tool in organisational processes and personal concerns. Venkatesh et al. (2003) find similar results in their model of technology use. In their work, performance expectancy, as the "extent to which an individual believes that using the system will help them improve their job performance", represents a relevant factor similar to the perceived usefulness of the Technology Acceptance Model (Davis 1989). Regarding the norm, the interview data also divides this factor into legally established norms and (organisationally) culturally determined aspects.

Proposition 1: For ML-based HRA, an understanding of related studies is not sufficient to explain individual adoption. Instead, we propose additional important determinants for PBC (data-based and algorithm-based efficacy), attitude (compatibility of tools and tasks, personal concerns, improving practices through innovation) and norm (organisational culture and legal framework).

Most ML characteristics have an influence on behavioural beliefs

Furthermore, our results suggest that several additional ML characteristics drive perceived algorithm-based efficacy as well as attitudes to the use of the displayed ML tool. First, to make ML-based HRA useful, predictors' diagnostic results – provided by the XAI visualisations – are an important facilitator in terms of coalescing employee turnover predictions (Chowdhury et al. 2022). Identifying additional uses of the HRA tool when understanding the causes of employee turnover suggests a more evidence-based management practice, which is theorised as an important enabler of HRA in an organisation (McCartney and Fu 2022, p. 29). The studies by Kim et al. (2023) and Haque et al. (2023) reveal that XAI visualisations in particular contribute to user understanding and adoption, when appropriately designed. To summarise, in most cases, higher transparency leads to a higher attitude and PBC. We thus provide an empirical example demonstrating that with sufficient ML transparency, the various burdens on users (emotional, mental, prejudices, etc.) (Park et al. 2021) can be overcome when introducing ML-based HRA. However, in line with other research (Schmidt et al. 2020), we also find that these effects can be reversed when the rational explanations of the model and the reasoning of experts contradict each other.

Second, our interviewees were reluctant to automate an entire decision, for example a promotion, by delegating it to the HRA tool, as most of them did not expect the respective tool to possess the necessary skills to solve the task adequately on its own. This finding is similar to the results by Dietvorst et al. (2018), who identified a significant aversion to fully automated predictive analytics tools that vanishes when participants get at least some degree of control over the underlying decision. Lee and Cha (2023) confirm this notion by showing that choosing augmentation over automation is one of the two key factors in adopting Artificial Intelligence recruitment systems.

Third, we observe a difference in the beliefs of the interviewees, who expect the displayed HRA tool to have self-learning capabilities or not. In sum, our empirical data indicates that interviewees are more forgiving of an error-prone prediction in the first case, as maybe because they expect the HRA tool to subsequently improve upon its past mistakes and thus increase algorithm-based efficacy. Note that this finding is in line with Reich et al. (2023) and Berger et al. (2020), both of whom identify self-learning capabilities as an important factor in mitigating algorithm aversion. Our study extends these interesting points by the fact that self-learning ability may be taken for granted by using the ML term, while such abilities (e.g., reinforcement learning) are not even implemented.

Fourth, as proposed by Vargas et al. (2018), our interviewees were interested in trying out the displayed HRA tool to assess its capabilities and then form an attitude on it. In addition, indicators of initial anxiety about the ML-based HRA tool's capabilities were apparent (see Table 7), which can be addressed with trialability. This finding is in line with the *Innovation Diffusion Theory* (Rogers 2003). In summary, we therefore propose:

Proposition 2: Several ML characteristics influence attitude and PBC in relation to the intention to adopt ML-based HRA: (a) the degree of transparency created, (b) the choice of automated usage, (c) the implementation of self-learning capabilities and (d) the enabling of trialability.

Consistent with Neumann et al. (2022), we find no reference in the interviews to ethical considerations regarding the adoption of the ML-based HRA tool in our empirical setting from public sector. Our results show that fairness and non-discrimination were not critically questioned, even when potential biases were highlighted by XAI visualisations and the interviewees were explicitly asked about them. This outcome is alarming, as HRA, and especially ML, can foster discrimination and create various risks for employees and the organisation (Tursunbayeva et al. 2022). For example, according to the General Data Protection Regulation of the European Union, the use of an ML model that includes protected class variables for individual decision-making can be considered a legal case of discrimination, referred to as “disparate treatment” (Goodman and Flaxman 2017). Lee and Cha (2023) confirm that solving the fairness problem remains complex, even if this complexity mitigates discovering an unfair decision basis.

Proposition 3: Fairness does not matter for the individual deciding whether to adopt ML-based HRA and must be ensured by other appropriate measures.

Implications for practice

In practice, organisations seeking to leverage the potential of efficiency gains through ML-based HRA might try to increase adoption at the individual level. Our results reveal that various adjustable modifiers exist during adoption, in particular the degree of automation in algorithm-based decision-making and provided transparency. Most importantly, technical measures (such as XAI) can positively or negatively influence both PBC and attitude, because providing understandable visualisation allows the user to compare the included predictors, as well as their effects with their intuition.

However, in trying to increase the intention to adopt ML-based HRA on an individual level, organisations should be careful to avoid pitfalls. For instance, negative examples have already demonstrated that biases can lead to unfair decisions based on ML (e.g. Alon-Barkat and Busuioc 2023). With the risk of resulting high social and economic damage, the consideration of ethical challenges is necessary in HRA projects (Langer and König 2023; Edwards et al. 2022, p. 5). As legislation for the responsible use of ML comes into force soon (AI HLEG EU 2019), organisations need to address such potential unfairness proactively. Thus, our finding, namely that ethical considerations and fairness of HRA in the early adoption stages were not challenged by the interviewees, is alarming and should therefore be paid careful attention in practice. Neumann et al. (2022) note that specifically the early adoption phases of ML applications are characterised by (1) a focus on positive business cases, (2) reliance on external partners, (3) change management processes to increase acceptance and (4) little to no real recognition of ethical considerations such as algorithm accountability and fairness. Thus, we specifically advise organisations to ensure proactively the inclusion of ethical considerations in the early stages of adoption and to implement internal policies and approval procedures with the help of internal or external expertise.

Ultimately, the responsible use of ML-based HRA can only be achieved when HR professionals have the knowledge necessary to evaluate ML models critically, based on the transparency provided by technical measures such as XAI visualisations and internal guidelines (Langer and König 2021). However, Vargas et al. (2018) note that HR professionals have low levels of quantitative self-efficacy (fear of maths/statistics, lack of quantitative training, low awareness of analytics, lack of resources and organisational support to promote analytics and its tools). Our results extend these findings, which suggest investing in training initiatives that demonstrate the importance of achieving ML transparency and in turn encourage the acquisition of skills specifically to interpret performance statistics of ML algorithms or XAI visualisations.

Limitations and further research

There are two main points that limit the findings of this study. First, our qualitative approach is based on the manual coding of interview transcripts; however, we took several measures to ensure the validity of our findings during the coding process and after the final analysis (see Figure 17). For example, the credibility of our findings was established by independent coding by two of the authors in three coding steps. In addition, our results were critically reflected on the basis of existing evidence (Vargas et al. 2018) and by the third author. The findings were verified by inter-coder and intra-coder reliability (Miles and Huberman 1994).

Second, while the methodological choice of a single case study has a solid foundation in HRM, and recent studies using this method advance the field considerably (e.g., Ellmer and Reichel 2021; van den Broek et al. 2021; Remneland Wikhamn et al. 2023), this methodological choice limits the transferability of our findings (Flyvbjerg 2006). Nonetheless, it offers the advantages of an in-depth investigation of HRM practices with a heterogeneous interviewee population (e.g., diverse backgrounds and experience) and an examination of deep cause-effect relationships (from ML characteristics to HRA adoption) that are overlooked in broader studies. In our particular case, the results thus pave the way for future quantitative studies that can explain the individual adoption of HRA more holistically and further develop the previous framework by Vargas et al. (2018), which can only explain about 35% of the observed variance. To achieve this, the exploratory and qualitative nature of our study leaves the following concrete possibilities for future research. First, future studies should examine multiple organisations to further validate the transferability of the three propositions for individual HRA adoption. Second, we invite future research to formulate and quantitatively test hypotheses based on our proposed qualitative model. Especially, the effects of the automated usage of ML predictions and ML transparency provide interesting opportunities in this regard, as they both affect PBC as well as attitude. In addition, our study is not able to provide insights into the effect strength of the assumed causal relationships between ML characteristics and the intention to adopt. Third, we focus on the first implementation of an ML-based HRA tool, which means that the key beliefs and experiences identified, as well as underlying ML characteristics, may not apply to a more mature stage of ML adoption. Therefore, given that information systems research has found a considerable number of factors influencing the intention to use HRA (e.g. Mahmud et al. 2022), future research could investigate whether the effects and significance of certain factors change over the course of the implementation and utilisation phase. For example, does the importance of ML transparency decline as users of ML-based HRA gain experience over time and learn that the system provides (in-)accurate results? Fourth, we agree with the widespread view that HRM systems need to be tailored to the individual case (e.g., Remneland Wikhamn et al. 2023), which is why we also call for more qualitative research examining the individual adoption of technological advances in ML-based HRA.

Conclusion

In contrast to existing technological tools, ML-based HRA generates unique challenges, most notably the potential opacity of the rationality used by models to formulate predictions, as well as the potential to automate HRM decision-making fully. This study provides deeper insights into behavioural beliefs determining the decision to adopt ML-based HRA from an individual perspective and sheds light on how ML characteristics affect it. Based on the focused interview methodology, we introduce novel propositions and an extended qualitative framework with new constructs of important factors from the perspective of end-users of individual HRA adoption. Investigating the lines of reasoning also reveals that potential ML model users do not include fairness considerations in their decision to neglect or adopt the tool. We hope our findings help to guide both the interdisciplinary research on HRA and organisations to a successful path in their mission to achieve the responsible proliferation of ML-based HRA.

3 Comprehensive discussion

This section provides a comprehensive discussion of the three studies presented in the last section. In particular, the overarching key findings for the in-depth case study are discussed along with other findings from the literature on algorithmic HRM and other high-risk decision-making applications from the public sector (see *Figure 3*). Recall the overall aim of this thesis, to help shift the use of AI in algorithmic HRM to the bright side of the continuum whilst mitigating the dark side. In particular, this section discusses how the results of all three studies can be interpreted holistically in terms of meeting the requirements for responsible AI (see *Figure 4*) and how the individual and organisational adoption of algorithmic HRM can be achieved. The comprehensive discussion in this section focuses on what can be learned from the three studies in relation to an important blind spot in algorithmic HRM (Lamers et al. 2024): The technologies surrounding ML and AI are often oversimplified using the black box metaphor so that "the importance of social contexts is not recognised in which technological artifacts are formulated, implemented, interpreted, and appropriated" (Kim et al. 2021, p. 234).

3.1 Mitigating the dark side of AI in high-risk decision-making

A first very important finding of all the studies is that the seven requirements for responsible AI cannot be considered and treated separately. Instead, transparency seems to be not only a requirement in itself, but also an important prerequisite for fairness, non-discrimination and diversity (see study 2) as well as oversight and human control (see study 3). However, transparency can sometimes interfere with privacy and lead to a conflict of interest. If important features of an ML model rely on protected personal data, opening up an opaque black box model simultaneously leads to more transparency and less privacy. This results in a trade-off which, like the trade-off between predictive performance and transparency (see study 1) or predictive performance and diversity (see study 2), needs to be addressed with a holistic optimisation problem in the ML development process. While the implications of this trade-off in the search for AI are not entirely new, technical studies (Krishna et al. 2023) have only recently begun to address it in relation to the "right to explanation" and the "right to be forgotten" (GDPR 2016). Overall, the trade-offs collectively illustrate that the complexity of responsible AI results from not only the fulfilment of seven demanding requirements alone, but also their interdependencies.

For example, the three studies also reveal interdependencies with the accountability requirement, in that the organisation's internal self-regulated guidelines address the accountability (also referred to as 'intervenability' and 'responsibility') to emphasise the need for flexibility to deactivate an AI model, as well as training responsible managers to decide on its use and serve as the first point of contact in case of criticism and scepticism about the system (Network AI in labour and social administration 2022). In the interviews in the third study, these guidelines are substantiated by the statement that HRM employees who use ML models as algorithmic decision-making aids are also responsible for critically questioning

the correctness and fairness of decisions. In this sense, the interviewees believe that it is not the ML model itself that must be responsible for decisions but the output of the socio-technical system consisting of algorithmic advice and human decision-makers. Complete automation is not desirable, though, as the human connection and interpersonal perspective are missing. The implications of this insight led to the motivation for study 2.[14] In summary, studies 2 and 3 show that in algorithmic HRM, the way ML models and predictions are perceived by HRM staff is critical to achieving accountability in ML-based decision-making, not only in terms of responsibility for monitoring a model, but perception also determines the ultimate impact of ML predictions on business process decisions. Furthermore, at least in this case study, an augmentation approach with a human decision-maker is preferable to automated approaches employed to fulfil not only the requirements of responsible AI in terms of fairness and human oversight, but also accountability. However, while the results on transparency and fairness of this case study may be applicable to automation through algorithmic HRM (e.g., see prescriptive analytics of gig platforms in the introduction section), this thesis leaves open questions on how the requirements accountability and human agency and oversight can be fulfilled. If algorithms are to make autonomous decisions, it is essential to at least establish a socio-technical supervisory body that rigorously assesses their impact on individuals on a regular basis (Charlwood and Guenole 2022; Kelan 2023; Edwards et al. 2022).

It should be noted that the case study also has limitations in terms of what can be learned about fulfilling the requirements of responsible AI. For example, in study 2 for diversity, non-discrimination and fairness, this work addresses important overarching open research questions on gender equality, such as "How can historical bias in HR datasets be mitigated?" (Kelan 2023). However, due to limitations in the respective datasets in this case study, equally important and more specific research questions regarding non-binary, queer and intersectional inclusion remain unanswered. As only gender-binary data are available, similar analyses of adverse impacts for these subgroups, such as between binary genders, with appropriate fairness measures (see study 2) are not possible. Future research should focus on how the above and other minority groups can be adequately accounted for in ML models. The results of this thesis suggest that ML offers new opportunities to overcome the limitations of traditional statistical methods used for inductive research methods or algorithmic HRM and HR analytics, due to its ability to model multifaceted and nonlinear relationships (see study 1). However, this potential can only be realised if the data are available for analysis, which requires the organisation to collect and process inclusive gender-specific information beyond binary assignment. The labelling and training of equally accurate models for all subgroups often result in unsolved problems in practice and leave room for future research (Kelan 2023). Interestingly, the question whether non-binary information regarding gender is

[14] Please note that the development of the framework in study 2 post-dates study 3.

collected and analysed is another example of the aforementioned trade-off between diversity and the need for data protection and privacy in responsible AI.

3.1.1 Understanding and addressing cognitive human biases

Documented examples of the dark side of AI and algorithmic HRM (see introduction) can be attributed to either biases in the data basis (e.g., historical hindsight), design biases in the ML development process (e.g., subjective opinions or wrong intuition of the developer) or the translation phase of predictions into actual decisions, due to human bias (Kelan, 2023). The technical frameworks proposed in studies 1 and 2 aim to mitigate the first two reasons (data and design) for negative consequences through statistical measures derived from a technical perspective from the results of the ML model. The interviews in study 3 point to possible cognitive human biases which, if not properly recognised and addressed, could undermine the effectiveness of the proposed technical frameworks in the later stages of ML output utilisation; for example, even if the ML predictions are accurate, fair and transparent, they could be misused by managers in day-to-day decision-making. The next section the builds on the interviews (see Table 7, Table 8, Table 9) and extends discussion of study 3 regarding cognitive human biases and their implications.

Documented examples of negative consequences in algorithmic HRM can usually be traced back to (1) distortions in the data foundation (e.g., historical hindsight), (2) design distortions in the ML development process (e.g., subjective opinions or wrong intuition of the developer) or (3) the translation phase of predictions owing to actual decisions resulting from human biases (Kelan 2023). The proposed technical frameworks in studies 1 and 2 aim to mitigate the first two reasons (data and design) through statistical measures derived from a technical perspective from the outputs of the ML model. The interviews in study 3 point to possible human cognitive biases. The discussion is based on the notion that the effect of technical measures will not lead to responsible AI unless all three reasons are addressed in conjunction. For instance, even if the ML predictions are correct, fair and transparent, they could still be misused by managers in everyday decision-making. The next section discusses a number of cognitive human biases and their implications.

The negative consequences of ML-based high-risk decisions, with real-world impacts stemming from cognitive human bias, have not been documented in algorithmic HRM but in the public sector. To create risk profiles of people applying for childcare benefits, for instance, the Dutch Tax and Customs Administration used algorithms in which 'foreign-sounding names' and 'dual nationality' were used as indicators of potential fraud. As a result, thousands of (racialised) low- and middle-income families were screened, falsely accused of fraud and asked to pay back benefits they had obtained completely legally. The algorithms thus led to legally punishable racial profiling (European Parliament 2022). Alon Barkat et al. have investigated the case further and point to biases such as over-reliance on algorithmic advice (‘automation bias’) or reliance on conformity to stereotypes (‘selective bias’). However, they find no evidence of this prejudice in laboratory experiments and argue that bureaucrats no longer rely on

discriminatory outcomes, due to increased awareness of discrimination and algorithmic bias following the scandal (Alon-Barkat and Busuioc 2023). In this context, in a field study that contrasts the problematic nature of responses to artificially created environments in experimental methodological settings, study 3 provides valuable views on potential automation and selective bias prior to the introduction of the systems. Interestingly, there is no evidence of bias towards automation, as the respondents highlight the possibility of altering the automated prediction through human judgment, as well as have an overall critical perception prior to the explanations, suggesting that people want to challenge the predictions. However, the results of the study point to possible human biases similar to the above 'selective bias' that could be problematic in ML-based high-risk decision-making, most notably confirmation bias. That is, the interviews in study 3 indicate overconfidence in the ML model, as respondents' attitudes towards the ML model and explanations increase as these explanations become more consistent with intuition. This notion may undermine the thesis that ML transparency provides an interface for correcting an ML model's incorrect decisions and important human oversight (see study 1). XAI research is aware of the problem of identifying or selecting explanations that are consistent with intuition, but studies rarely discuss how to mitigate this effect. Ha and Kim (2023) are a recent exception in this regard and prove that the use of textual explanations, in combination with the provision of a priori information, can strengthen trust and at the same time mitigate cognitive biases. This a priori information, which refers to general knowledge that is known prior to the application of ML, can come for example from domain knowledge or scientific theory. This idea is also used in the sociotechnical approach, and an empirical demonstration of the framework in study 2 supports the idea of including ML assessments with a priori information. Textual explanations extracted via post-hoc XAI methods were not applied here, but they could be implemented in the interface, which in turn could further improve the quality, fairness and value of the sociotechnical output. When adapting the sociotechnical framework of study 2, it is therefore advisable to select the XAI method based on the requirements of XAI stakeholders. For example, Riveiro and Thill (2021) find that the selection of XAI methods can partially fulfil the expectations of the addressees of explanations by answering the questions of causation ("What are the reasons for this prediction?") or counterfactual reasoning ("What is the reason why it is not this prediction?"), depending on the requirement.

Another hitherto unexpected type of cognitive bias reported in study 3 is based on the hype surrounding generative AI – and ChatGPT in particular (see 1.1 Motivation). Some interviewees initially showed high confidence in the ML models' predictions and pointed to the self-learning capabilities of particular example, using viral AI systems such as ChatGPT as an analogy (e.g., "*So, if I understand machine learning correctly, it [the ML model used to predict voluntary turnover] is a self-learning system, and the more often I run it through, the better my predictions will be*"; "*I think scepticism is healthy, even towards Chat GPT, but the machines are constantly learning*"). Interestingly, neither self-learning capabilities are part of the description of the developed ML model nor was the iterative development of

models, as described in the sociotechnical framework in *Figure 14* in study 2, communicated. Instead, the ML model, which is trained on the currently static dataset with a fixed process (see *Figure 7* in study 1), is only updated when new or more recent data are introduced into the ML model. In short, interviewees assumed that self-learning capabilities are implemented because of the term "ML" and its close conceptual or terminological association with AI technology. These incorrect assumptions, implying an hybrid form of association fallacy and technological optimism bias (Clark et al. 2016), might lead to overconfidence in the ML model's accuracy, even if the predictions are incorrect. This result supports the literature, which shows a significant impact of terminology, meaning simply terming or naming algorithmic tools differently (e.g., ML, AI, model, algorithm, computer programme, robot) (Langer et al. 2022). For instance, in the specific case of HR, the authors point out that the term 'AI' is perceived as less beneficial, which could lead to a stronger public outcry in contrast to terms such as 'algorithms in HRM'. While the impact of naming the tool is somewhat surprising (and not reflected in the motivation behind or hypotheses in the three studies), they nevertheless have important implications for practitioners looking to adopt algorithmic HRM and ML in general. Individual-level adoption and appropriately calibrated confidence leading to responsible use can be strategically influenced by naming the tool precisely according to its capabilities. In addition, the provided information about the model needs to explain the capabilities of the algorithmic model in more detail, as the terminology could be interpreted differently by various stakeholders. In addition, this finding supports research suggesting that generative AI accelerates and drives the adoption of a data-driven decision paradigm and culture (Davenport and Bean 2024), which in turn is also a key enabler for narrow algorithmic HRM such as the predictive algorithms used in this case study.

Finally, this section discusses the implications of possible wrong decisions resulting from human cognitive biases. The main question that arises here is: Who is accountable if an ML-based decision leads to negative consequences? Respondents in study 3 do not directly attribute accountability, e.g., in terms of non-discrimination, to the ML model and its predictions as proposed in the EU AI Act (see *Figure 3*). This means that developers and data analysts may not be held accountable for their developments, as respondents believe that the models, as well as the data used, adequately reflect the real world (see study 3). Instead, according to the interviewees, end users (e.g., HR managers) and other stakeholders in employee turnover predictions are accountable for recognising incorrect results and acting accordingly, i.e., only using them when they think the predictions are fair. Regardless of the intended HRM application, respondents are interested in the overall performance of the system, which ultimately has an impact on employees. In light of the human cognitive biases above, this may be problematic; moreover, HRM staff may lack background knowledge, e.g., about the risk of transferring biases in data to ML models. For example, it is worrying that the respondents in study 3 are not aware of the problematic consequences of using protected group information as an input feature. Study 2 therefore suggests that accountability should be distributed across the entire process of developing and

applying ML. The proposed framework specifically targets fairness but could be extended to other requirements such as accountability. Intuitively, technical specialists developing ML should be responsible for the predictions, as they have the necessary technical knowledge to recognise biases in data and ML predictions. However, given the cognitive biases mentioned above, it seems clear that they cannot be responsible for the overall decision. This notion is highlighted in the fairness assessment framework proposed in study 2, as sociotechnical systems theory suggests a holistic approach to optimising the overall system. Consequently, technical and HRM staff are jointly accountable for overall decisions, with the technical staff being more responsible for the reasonableness of the model's predictions, and the human resources staff being more responsible for the final judgment. The establishment of clear processes and organisational structures in organisations is particularly important. This holistically distributed accountability of a sociotechnical system is specifically important because Fest et al. (2023) find evidence that data professionals may exert discretion and are aware of public values, but their value-sensitivity often does not translate into responsible practices. Instead, they use a variety of arguments to dissociate themselves from, or downplay, their responsibilities. Overall, the studies show that without establishing clear guidelines for accountability, the promise of a more objective basis for decision-making is undermined and instead leads to more discrimination and injustice (Alon-Barkat and Busuioc 2023).

For the HRM function, this means that they are primarily accountable for the algorithmic HRM tasks that fall within the core area of their professional background knowledge, albeit they also require knowledge of the technical fundamentals in order to appropriately interpret the data, models and explanations (e.g., quantify uncertainty statistically). This applies in particular to post-hoc explanation methods, which, as study 1 argues, may produce unreliable results and therefore require suitable background knowledge for adequate use. Due to potential unreliable results, the literature is divided on its application to high-risk decisions. In this regard, the contribution and implications of this case study to this debate are discussed in the next section.

3.1.2 The importance of XAI techniques, and their actual contribution to responsible AI

The in-depth case study above provides rare and valuable insights into the contribution of XAI in terms of fulfilling the requirements of responsible AI, as the three studies uniquely provide a complementary assessment of the necessity and implications of post-hoc XAI methods. This section initially briefly reviews the first study's discussion on which two dogmatisms emerged in the literature, and then it extends this discussion by addressing the appropriateness of post-hoc XAI methods in achieving ML transparency by adding the aspects of effects on fairness (study 2) and adoption (study 3). Following the motivation behind and introduction to this thesis, the overarching question uniquely answered by the collection of all three papers is: How can post-hoc XAI methods help to shift the dark side of AI to the bright side?

Under the first dogmatism, researchers from various disciplines call for avoiding complex ML algorithms in combination with post-hoc explanatory methods for high-risk decisions and instead suggest using simpler algorithms for transparency-by-design approaches (e.g., Rudin 2019; Ghassemi et al. 2021; Vale et al. 2022). The reason for this is that, due to the approximate nature of post-hoc XAI methods, these explanations may be inaccurate, unstable and therefore potentially misleading. Under the second dogmatism, research in fields like interpretable machine learning or XAI proposes the advantages of post-hoc explanatory methods to utilise complex ML algorithms in decision-making, thus maintaining high predictive performance while achieving sufficient comprehensibility (e.g., Arrieta et al. 2020; Langer and König 2023; Chowdhury et al. 2022). While the first study presents a nuanced view of when the latter post-hoc XAI approach is justified in algorithmic HRM applications, depending on the complexity of the data-generating business process (e.g., voluntary employee turnover prediction), papers two and three extend this idea by assessing the impact of post-hoc explanations.

In the sociotechnical framework of the second paper, ML transparency is the key element in creating a human-ML interface that establishes a link between human expertise in HRM-related assessments and ML strengths in predicting complex phenomena. The empirical demonstration part of the design science methodology showed that XAI methods provide sufficient transparency for the sociotechnical approach to ML fairness assessment. First, the HRM experts were able to understand the reasons for a high or low predicted turnover probability on an individual level. Second, post-hoc explanations allowed them to understand from a different context (post-pregnancy integration, etc.) the predictor effects in relation to measures to prevent employee turnover. Third, based on this information, the HRM experts were able to identify spurious predictors potentially leading to unfair bias and thus incorporate their perception of fairness into the next iteration of the technical ML development process. This finally led to models with fewer adverse impacts on certain protected groups (fulfilling the 4/5- rule), thereby fostering the diversity, non-discrimination and fairness requirement of responsible AI in algorithmic HRM. Both the promise of fairness in algorithmic HRM achievable through ML transparency (e.g., Leicht-Deobald et al. 2019; Gal et al. 2020; Köchling and Wehner 2020; Choudhury et al. 2021; Cheng and Hackett 2021; Langer and König 2023; Meijerink et al. 2021) and the promise that post-hoc XAI methods achieve sufficient ML transparency (e.g., Chowdhury et al. 2022; Haque et al. 2023) are thus realised and valid in this case study. However, study 2 illustrates that the ML transparency of post-hoc XAI methods necessarily leads to fairness. Instead, we found no evidence that differential treatment and adverse impacts between protected groups (e.g., women and men) lead to a request to exclude protected features because it could affect the predictive performance of the ML model or does not reflect the real world (note that the data that is reflecting the real world could be biased), or to a lower intention adopt the system. This is why study 2 extensively addresses the notion that ML transparency from post-hoc XAI methods needs to be combined with automated technical approaches for ML bias mitigation, in order to create interconnected processes for mutual and iterative processes to optimise the holistic sociotechnical

system. This example illustrates that post-hoc XAI methods – and ML transparency more generally – are an effective building block, but not an all-encompassing solution to meet the diversity, non-discrimination and fairness requirements of a specific responsible AI implementation.

Further, looking at impacts on the individual adoption level, the interviewees in study 3 revealed increased transparency by the provisioning of post-hoc XAI explanations leading to higher PBC and attitudes – and thus to a higher intention to adopt the ML model in a variety of tasks and HRM processes. For example, after examining the post-hoc XAI explanations, more potential applications of the ML model were revealed in various HRM processes (in study 3) and actionable measures identified (e.g., careful attention and reactions to long-term absenteeism and reintegration). This result is in line with studies demonstrating that with sufficient ML transparency, the various burdens on users (emotional, mental, prejudices, etc.) can be overcome when introducing ML systems (Park et al. 2021). Interestingly, there are some exceptions in study 3, whereby respondents showed less PBC and intention to adopt the ML model after seeing that the explanatory visualisations provided did not include certain information (e.g., the percentage of hours worked from home) in the ML model predictor, which he argued was an important factor in predicting turnover. Similarly, the attitude and intention to adopt the ML model decreases if the predictor's effects do not match one's own intuition. As a result, the information provided by the XAI led to a lower intention to adopt the ML model, which seems counterintuitive. Nonetheless, this reaction is consistent with research showing that when explanations are not consistent with the causal reasoning of experts (e.g., when models are perceived as too simple and use only part of the relevant information as an input), increasing transparency leads to a decrease in trust and acceptance, which is contrary to the common narrative mentioned above (Schmidt et al. 2020). This effect may especially be a relevant factor in algorithmic HRM due to its historical qualitative nature, depending on HRM experts' judgements. First evidence arose from studies that show a substantial difference between ML models' correlations used for predictions and experts' causal knowledge, specifically in the algorithmic HRM domain, which sometimes can explain why HR practitioners avoid using ML models for tasks like employee turnover prediction despite their good performances (Meddeb et al. 2022).

This is especially interesting when considering the findings regarding confirmation bias in the last section. On the one hand, the case study shows that ML transparency can indeed increase or decrease the trust in and acceptance of ML models, so that, in principle, an inappropriate level of reliance leading to under- or overconfidence in predictions might be mitigated (Glikson and Woolley 2020). On the other hand, the direction of calibration (decreasing or increasing confidence) is not always necessarily directed towards the desirable sweet spot, in which incorrect predictions are identified and corrected accordingly and correct predictions are used accordingly. Instead, the direction is determined by the agreement of the prediction with its explanatory predictor effects with one's own intuition. This has important implications. First, if the initial intention to adopt an ML model is high, the user might choose elements of XAI explanations that are consistent with intuition, potentially leading to even more confidence and

thus overconfidence, and vice versa. Second, the fundamental promise of ML to gain knowledge inductively from data (see study 1, inductive research method) can be neglected if the newly discovered patterns do not match common understanding and previous intuition. In conclusion, the adoption and actual use of the ML model can be fostered by transparency provided by post-hoc XAI methods. However, post-hoc XAI explanations cannot solve all human cognitive biases and leave important questions open, such as what should be explained (and how) in order to increase confidence and influence behaviours such as individual adoption (Shin 2021). One interesting research attempt to solve these issues is the relatively new field of causal ML that has emerged in the last five years from AI and XAI in general (Kaddour et al. 2022). In this case study, extracting causal rather than correlational relationships through ML could help HRM experts explore even more new ways of drawing conclusions from ML that lead to predicting employee turnover, which also could foster the development of theories with an ML-based inductive method (Cheng and Hackett 2021).

Returning to the overarching question of this section, namely for what purpose should post-hoc XAI methods be used in high-risk decision-making, the case study shows that the transparency achieved is helpful for technical ML developers (i.e. data scientists) to (1) debug and refine models (e.g., to identify spurious predictors), (2) extract further insights inductively, to find new insights and discover new knowledge from data (see study 3), and (3) contribute to the fairness assessment of ML models (see study 2). For findings one (e.g., Rudin 2019) and two (e.g., Choudhury et al. 2021), there appears to be broad consensus in the management and technical literature. For the third finding, however, there is an ongoing debate, with some literature on transparency in algorithmic HRM suggesting the fair use of ML solely through the application of post-hoc XAI methods (Chowdhury et al. 2022). Extending the discussion of the first study, this case study highlights that post-hoc explanations for high-risk decision contexts are not only potentially misleading or due to the technical approximation, but they are also due to potential misinterpretation by human cognitive biases (see results study 3 and discussion above). How to mitigate the misapplication of post-hoc XAI methods and establish a critical-rational view of outcomes therefore appears to be an important advance in the algorithmic HRM literature (reflecting the specific research questions raised by Langer and König 2023). Thus, by providing a realistic empirical example that goes beyond the reluctant adaptation of technologies in algorithmic HRM, this work contributes to the claims of the algorithmic HRM literature highlighting the effect of XAI in fairness assessment (Rottman et al. 2023, p. 1440), ML development with HRM involvement in practice (Charlwood and Guenole 2022, p. 737) and understanding important predictors of HRM phenomena in a scientific inductive methodology (Yuan et al. 2021, p. 16).

3.2 Realising the bright side of algorithmic HRM

3.2.1 Value contribution of algorithmic HRM, and determinants for successful AI adoption

This section discusses whether the high value proposition (see the 'bright side' in the introduction section) of algorithmic HRM and AI technology in general is actually realised in the case study. This is worth exploring further, as recent reviews on algorithmic HRM adoption reveal that companies rarely take advantage of AI's potential for more accurate, objective, and effective decision-making, despite significant investments in the technology (Wang et al. 2024). One reason for this is found by recent studies explaining the phenomenon of AI resistance at an individual employee level (Golgeci et al., 2025). In this context, the in-depth case examination can contribute to our understanding and current discussion about the determinants of successful AI adoption.

This section explores this aforementioned aspect and discusses the implications of the specific ML model in this in-depth case study in terms of (1) the challenges of status quo approaches to employee retention, (2) the added value of the ML model for predicting employee turnover, and (3) the determinants for the successful adoption of algorithmic HRM.

(1) Challenges of the status quo approaches for employee retention

The organisation's existing approaches to combating employee turnover are mostly based on survey data from annual individual employee development meetings and documentation relating to off-boarding processes. However, the status quo approaches have the limitation of being retrospective and does not allow for proactive action to retain employees (Rombaut and Guerri 2018; Yuan et al. 2021). In addition, these survey-based methods help measure the complex reasons for employee turnover, but they can be skewed by potential social factors or, more commonly, employees respond to what they believe will benefit them the most (Gieter et al. 2012; Yuan et al. 2021). Further, these methods are often limited because they target employee turnover *intention* (Lum et al. 1998; Gieter et al. 2012; Holzwarth et al. 2021) rather than gathering timely longitudinal data and analysing actual turnover. This is not always appropriate because, as examples from the public sector show for certain occupations (e.g., teachers), the link between turnover intention and actual turnover is relatively weak (Grissom et al. 2016, p. 242).

(2) The value-added contribution of the ML model

As documented in the confusion matrix of the test data (see *Table 5* in study 1), the ML model is indeed able to predict half of the cases of voluntary employee turnover. In addition, when these results were presented to HRM experts in study 3, most of them agreed that the predictive performance of the ML model is useful for a number of HRM applications despite its imperfections (see study 3). Interestingly, there are potential areas of application for turnover forecasts, not only within HRM processes primarily focused on employee retention and workforce planning, but also for all other core HRM processes, including recruitment and selection, development and career development and performance measurement. In addition, the interviewees noted that the model and XAI explanations generally allows

to gain new and relevant insights into the identification of personas, subgroups or profiles at high risk of turnover, as well as the original reasons and causes thereof. This is an example of how ML-based inductive investigation of HRM phenomena benefits from not requiring prior assumptions and explicit hypotheses, leading to various opportunities to gain unexpected and innovative insight (Putka et al. 2018; Cheng and Hackett 2021). In summary, practitioners benefit from algorithmic HRM when it is used in addition to survey approaches to gain insights, including nonlinear relationships and interactions between predictors (see Study 1). This should also apply to other HRM processes beyond the context of employee turnover. The case study thus provides empirical evidence that algorithmic HRM can indeed be valuable for domain experts due to its potential to discover *new*, *objective* and *efficient* knowledge (van den Broek et al. 2021). To exploit this potential, it is important for practitioners to establish a bidirectional exchange of information between models and experts (see *Figure 14* in study 2). Based on shared information, a mutual learning process that evolves into a hybrid human-ML practice fosters discovered knowledge that is relevant and useful for HRM professionals. When this is achieved, according to this empirical example, the widely prevailing assumption is indeed that the knowledge discovered through ML adds value either by improving existing HR processes or by providing innovative solutions to key business challenges and strategic problems (Marler and Boudreau 2017; Wang et al. 2024).

(3) The determinants of successful algorithmic HRM adoption

As mentioned in study 3, a number of organisation-level factors are cited in the literature as examples of successful implementation of algorithmic HRM (Chowdhury et al. 2022; Budhwar et al. 2022; Malik et al. 2023), but there is less research on the individual level of algorithmic HRM adoption (Vargas et al. 2018). Here, the three studies (particularly study 3) in this thesis provide valuable insights into the reasons why employees (individual level) choose to use algorithmic HRM outputs and methods in their daily work. This is an important point to understand, as the choice made by individuals about the freedom to use a technology, from minimal use to collaborative use to the selective use of certain features, significantly influences the effectiveness of algorithmic HRM adoption (Wang et al. 2024). Study 3 has already shown that certain ML characteristics (transparency, automation, self-learning capabilities and trialability) raise the intention to adopt ML. Most importantly, recall that most of the potential applications of the ML-based employee turnover prediction model identified by the interviewees are explored after the post-hoc explanations have been provided, which again shows that the importance of ML transparency is a detrimental determinant for the adoption of algorithmic HRM. In light of the results of the study 1, XAI methods might be a correct choice for algorithmic HRM applications with high complexity and multi-layered phenomena. Choosing a more transparent ML model with lower predictive performance could lead to lower algorithm-based effectiveness and thus lower one's intention to adopt a model (see *Figure 9* in study 3). For example, Schmidt et al. find that when using simple ML models, the effect of more transparency could actually lead to lower trust in the

model. These complex relationships between design decisions during ML development and the eventual adoption of ML by end-users require stakeholder engagement in the early stages of adoption, before final implementation, in order to prioritise outcomes in light of the trade-offs mentioned in this thesis.

From a practical perspective, it is worth considering whether the value of algorithmic HRM still exceeds the costs, especially given the extensive technological and social efforts required to meet fairness and other requirements (e.g., see Figure 14), as well as reducing barriers to the successful adoption of algorithmic HRM. As this case study illustrates, the costs of meeting requirements for responsible AI may not be entirely separable from those associated with achieving algorithmic HRM adoption. For example, increasing ML transparency is both a requirement for the responsible use of AI and a factor that enhances individual adoption (see Study 3). However, the significant investment a company must make to develop an ML model and use it responsibly may lead to different prioritisations, which are further explored in the next section.

3.2.2 Reconciling economic value and responsible use with management accounting tools

This section is motivated by one of the central statement already mentioned in the introductory section: *Companies may prioritise economic goals over the responsible use of algorithmic HRM* (Charlwood and Guenole 2022). In adopting ML, organisations focus on positive business cases rather than taking a virtue-based approach that would benefit employees and business alike (Neumann et al. 2022). Lamers et al. (2024, p. 9) refer to this assumption as the "*algo economicus*"[15], which states that "HRM algorithms are non-human entities with the capacity to maximise utility and profit". That being said, the authors stress that the existing algorithmic HRM literature bases its assumptions on *algo economicus* and thus disregards the examination of compatibility with goals that may also be pursued, for example, by human actors shaping algorithms such as employee well-being.

This section therefore aims to move beyond the *algo economicus* assumption by exploring the compatibility of the two overarching goals of contributing economic value and meeting the requirements of responsible AI (as a basic prerequisite for algorithms that contribute to employee well-being). The underlying idea is to provide managers with a guide on how to identify application areas of algorithmic HRM where positive business value and the responsible use of algorithmic HRM can be reconciled, which in turn will help to realise the promise of the bright side of AI (see the introduction section). The following discusses in more detail (1) how the contribution to economic value can be quantified, using the example of the ML model in the case study, (2) the extent to which economic value and responsible use conflict and are compatible and (3) how algorithmic HRM applications that may contribute to economic value under responsibility requirements can be identified.

[15] 'Algo economicus' refers to homo economicus, an immoral individual utility maximiser who only engages in transactions with others to achieve his or her personal economic goals.

(1) Formulating the economic value contribution of algorithmic HRM

In order to understand the compatibility and contradiction between maximising economic value and the responsible use of ML, it is first important to know how the economic value contribution of algorithmic HRM manifests itself. The following example aims to quantify it and mathematically explain the impact of the decision boundaries resulting from fulfilling the requirements for responsible use of ML. Please note that the example is highly simplified, as the qualitative value contributions of ML (e.g., knowledge discovery) are not included because they are difficult, or even impossible, to quantify. Instead, the following approach focuses on the quantitative impact of the ML model in the case study on the organisation when operationalised to pre-select employees based on their turnover risk, with the aim of initiating countermeasures to prevent employee turnover (e.g., incentives such as salary increases for individual employees to reduce the risk of turnover).

A particular challenge caused by the uncertainty of ML models in this analysis is the aggregation of the value contribution of individual predictions used as decision support, as the impact of decisions varies at the individual level (for each employee) in terms of uncertainty (predicted turnover probability and statistical confidence) and benefits (avoided costs by retaining employees classified as 'true positive'). Further, the direct costs of countermeasures (e.g., incentive strategies for positive predictions, meaning the employee will leave within the next 6 months) must be considered. Therefore, in order to consolidate the value contribution made by individual-level activities, a "decision analytic thinking" method is needed (Provost and Fawcett 2013). The approach uses expected value as a statistical foundation to combine the confusion matrix (see *Table 5* in study 1) with a cost-benefit table of asymmetric and individual benefits and costs. On a consolidated organisational level, the value contribution $\boldsymbol{V}$ is calculated by comparing the benefit $\boldsymbol{B}$ for true-positive prediction $\boldsymbol{S}_{TP}$ with cost $\boldsymbol{C}$ for false-positive predictions $\boldsymbol{S}_{FN}$.[16]

$$\boldsymbol{V} = \sum_{i \in \boldsymbol{S}_{TP}} \boldsymbol{B}_{TP}^{i} - \sum_{i \in \boldsymbol{S}_{FP}} \boldsymbol{C}_{FN}^{i}$$

On an individual level, the benefit of true-positive predictions $\boldsymbol{B}_{TP}^{i}$ might be quantified by avoided costs through fruitful decision-support minus costs of measures triggered by positive predictions. In this case study, this corresponds to the cost avoidance of employee turnover $\boldsymbol{v}_{TP}^{i}$ (off-boarding, finding new employees, on-boarding process, lost productivity in the replacement time, etc.) minus the cost of incentives used to prevent turnover $\boldsymbol{c}$. It should be noted that the cost avoidance of preventing employee turnover $\boldsymbol{v}_{TP}^{i}$ varies between each employee i (e.g., based on seniority and skillset). Decision analytic thinking also takes into account the varying effects of incentives, in this case to reduce employee turnover risks. For example, identifying employees as 'true positives' may not add value if the turnover

[16] In this particular case, the cash-flow relevant costs for false-negative predictions $\boldsymbol{C}$*(FN)* and the benefits of true-negative predictions $\boldsymbol{B}$*(TN)* are zero, as no countermeasures are taken in these cases. However, it should be noted that practitioners might consider the opportunity cost of $\boldsymbol{C}$*(FN)* when comparing different ML models or other HRM approaches to retaining employees.

causes are unavoidable and therefore employee retention cannot be encouraged through incentives. Therefore, the value contribution of the model is calculated by the probability difference between the conditional probability that employee *i* stays after being targeted with incentives $\boldsymbol{p}(i_{stays} \mid i_{targeted})$ compared to the conditional probability that employee stays x when not targeted with incentives $\boldsymbol{p}(i_{stays} \mid i_{not\ targeted})$. Hence, the value contribution on the organisational level can be calculated as:

$$\boldsymbol{V} = \sum_{i \in \boldsymbol{S_{TP}}} [\, \boldsymbol{p}(i_{stays} \mid i_{targeted}) - \boldsymbol{p}(i_{stays} \mid i_{not\ targeted}) \,] * \boldsymbol{v}^{i} \; - \sum_{i \in \boldsymbol{S_{FP \cup TP}}} \boldsymbol{c}$$

Based on this formula, companies may selectively optimise $\boldsymbol{V}$ by picking employees that provide a positive value contribution. This means that the employee receives an incentive if:

$$[\, \boldsymbol{p}(i_{stays} \mid i_{targeted}) - \boldsymbol{p}(i_{stays} \mid i_{not\ targeted}) \,] * \boldsymbol{v}^{i} \; - \boldsymbol{c} \; > 0$$

This shows that in order to maximise the economic value contribution, only those employees might be addressed for whom the value of retaining them, weighted by the difference in the probability of turnover before and after the incentives, exceeds the costs of the incentives.

(2) Conflicts and compatibility of responsible use and the economic value of ML

An important implication of the above formulation of the economic value contribution is that employees who already have management responsibility due to their seniority, or who are more difficult to replace due to their special skillset (high $\boldsymbol{v}^{i}$), are more likely to receive even more incentives. As a consequence, inequality may increase further. This problem is neither addressed nor solved by using fairness or discriminations constraints like adverse impact ratios, as suggested in study 2.

In addition, the formulation reveals that the value contribution of an ML model depends directly on the accuracy of the prediction probabilities of turnover. To reduce the cost of ineffective incentives, the proportion of employees identified that fulfil an individual value contribution needs to be as accurate as possible so that, for example, fewer employees receive incentives who would not leave the company in either case. First of all, this implies that the technical trade-offs uncovered and explored in study 1 (predictive performance vs. transparency) and study 2 (predictive performance vs. diversity, non-discrimination and diversity) are directly translated into a trade-off between economic value proposition and responsible AI. These trade-offs thus directly imply the responsible use and economic value of AI, and thus tension and a conflict of interests may exist for an organisation.

While the above analysis reveals that reconciling the economic value proposition and responsible use of algorithmic HRM is a challenge and comes with trade-offs, it should be clear that companies should prioritise the latter; otherwise, there is a risk of legal violations when laws such as the EU AI Act come into force in the coming years. Furthermore, research has shown that ML could also have a negative impact on economic value contribution and instead lead to value destruction (Canhoto and Clear 2020). With algorithmic HRM in particular, negative impacts on corporate governance and social reputation is

a possible consequence – as negative examples from Amazon and other companies illustrate (Meyer 2018; Dastin 2022). In summary, conflicting priorities need to be managed when choosing how to initiate, implement or operationalise AI and ML applications categorised as high-risk (see *Figure 3*).

(3) Conflicting priorities in AI and implications for practitioners

The following section provides practical recommendations on how to manage the conflicting priorities of responsible use and the value contribution of algorithmic HRM and other high-stakes decision-making applications of ML and AI. In order to approach tension from the perspective of an optimisation problem, the specific rules for the responsible use of ML could first be translated into decision boundaries. The mathematical optimisation of ML functions can then maximise the value proposition within a solution space in which responsible use is already warranted. For example, the second study sets a decision boundary whereby the criteria of diversity, non-discrimination and fairness of adverse impacts (recall: adverse impacts are differences between protected groups) are set to a minimum of 80%. This example can be interpreted as a decision boundary that restricts the mathematical solution space of possible functions provided by ML from different algorithms and model development iterations. This general idea of establishing "ethical key indicators" provides a way to operationalise responsibility and create a foundation that can be audited, debated and improved (Lee et al. 2021, p. 542).

In addition, besides the obvious direct costs of additional steps for the implementation and responsible use of algorithmic HRM (activity and process costs: additional development of XAI post-hoc explanation methods in study 1 or the fairness assessment in study 2), there are indirect costs resulting from the potentially lower predictive performance of ML models fulfilling responsibility requirements compared to ML models that do not consider responsibility requirements. This should be considered when calculating the expected economic benefits of ML applications for high-risk decision-making applications. Moreover, it is essential for practitioners to consider and incorporate these direct and indirect impacts of responsibility requirements on value contribution when identifying algorithmic HRM applications that deliver economically positive business cases. That is the case because it can be assumed that companies will only implement algorithmic HRM in a responsible way that benefits both employees and the organisation and makes decision-making more accurate, objective and efficient if the business case still delivers a value proposition. Identifying the right applications of algorithmic HRM is therefore critical to realising the bright potential of ML and AI. To achieve this, three steps are required.

First, it should be recognised that the quantifiable business value of ML is not derived directly from predictive performance. Rather, business value is generated by the application and improvements in management decisions based on predictions from ML, or efficacy gains resulting from automation (Provost and Fawcett 2013). Quantifying business value therefore requires knowledge of the benefits and costs of the associated decision-making process and management accounting experience, rather than sophisticated mathematical knowledge or programming skills, and therefore it cannot be part of the data science role. Management accountants, specifically those specialised in the domain as HR controlling

experts, can play a crucial role in cost-benefit analysis for algorithmic HRM, as they have extensive knowledge of the business model, information about processes and costs and experience in providing objective decision support to management that is not fully available in other departments (Pickard and Cokins 2015). Likewise, the literature suggests that management accountants are able to "evaluate the effectiveness and efficiency of enhancements through ML-based technologies and prepare cost-benefit calculations" (Leitner-Hanetseder et al. 2021, pp. 551–552). Since the crucial information is shared between professional groups, HRM professionals and management accountants can only find valuable use cases for AI in collaboration.

Second, the literature points out that it is very important to define and establish easy-to-understand business metrics that form the basis for prioritising resource allocation as efficiently as possible (Wang et al. 2024). These metrics can include qualitative and quantitative measures of business process efficiency that make the value contribution of ML decision support transparent. These metrics may also be applied in the ML development phase. Instead of using out-of-the-box statistical metrics (e.g., Cohens Kappa in study 1), customisation and individual statistical measures more accurately guide the training steps of algorithm selection and hyperparameter tuning to ensure more realistic business impact optimisation. The use of custom metrics is particularly helpful in ensuring a positive value contribution in cases where the business impact of false-negative and false-positive is asymmetric, so that a cost-sensitive optimisation criterion (e.g., a pinball loss function) directly informs ML algorithms with real-world impacts, thus helping to optimise holistic output and decision efficiency. In algorithmic HRM, a cost-sensitive approach has already been successfully applied by Lawrance et al. (2021) when predicting employee absenteeism. In this case study, it is assumed that the cost of false-negative (an employee is classified as leaving the company, even though they will not actually leave) is lower on a qualitative basis than the cost of false-positive (an employee is classified as not leaving the company, even though they will actually leave), which could be further quantified for a customised optimisation criterion.

Third, it is critical that algorithmic HRM applications are aligned with and contribute to strategic business and HR objectives, which is not always the case in practice (Rasmussen and Ulrich 2015).

In conclusion, this section shows economic and ethical tension between the responsible use of ML and its contribution to economic value. To move beyond *algo economicus*, selecting fruitful applications is key to reconciling these conflicting goals. Applying management accounting tools and formulating a suitable optimization problem with appropriate decision boundaries in terms of responsible use of ML and AI can help practitioners in this task.

3.3 Outlook: Implications for the adoption of (generative) AI in HRM

screening, recruitment and selection, onboarding and open job description advertisements. Yet, it is impossible to say whether these impacts will have more positive or negative consequences on employees or organisations (Budhwar et al. 2023, pp. 615–616). This section aims to discuss implications for

current developments in AI, particularly generative AI, in light of the findings from the case study in this thesis, and to identify possible future directions for research and practice. The introductory section of this thesis explains the far-reaching implications of generative AI for individuals, organisations and society, as these latest ML-based services and products stand out due to the growing potential of increasingly larger underlying models with rapidly accelerating diffusion. In a recent *Human Resource Management Journal* publication, some of the leading researchers in the field suggest that the impact of these more general AI systems on organisations and employees will certainly be even more far-reaching than that of narrow ML models used in state-of-the-art algorithmic HRM trained on a specific dataset and for a clearly defined purpose (e.g., here: Predicting voluntary employee turnover in this specific case study). Instead, through generative AI, multiple tasks may be substituted, including currently central tasks and HRM processes such as job market

Interestingly, some researchers suggest a way forward for generative AI that corresponds to findings in terms of narrow AI, such as the call for (1) transparency including the application of XAI methods and (2) a sociotechnical theory approach (Budhwar et al. 2023, p. 615 and 619). Therefore, it appears that the results of the three studies included herein could make a significant contribution to the responsible and successful introduction of generative AI. While this may be true to a certain extent, there are some technical and managerial differences between narrow ML models focused on one aspect of a decision-making process and holistically applicable generative AI that limit the generalisability of the results.

First, a very important technical limitation may arise from the lack of sufficient XAI methods. In all three included studies, the post-hoc explanatory methods, namely SHAP, ALE and PFI, are of critical importance. While these methods are applicable to ML models that classify data based on structured tabular datasets, they are not compatible with large foundation models used in generative AI. To date, the computer science literature has suggested explanatory methods aimed at deep learning algorithms following a particular architecture (e.g., embedding explainable layers between multiple processing steps). However, whether these methods provide valuable explanations that enable people to answer questions such as "Why is recommendation X and not Y?" or "What should happen for the recommendation to change?" in response to the increasing complexity of deep learning architectures as well as sheer size of the models remains an open question. So far, attempts to make the most complex models more transparent by using simpler models to explain the reasons for the results of the complex models have only led to inadequate explanations without achieving an up-to-date breakthrough (Griffin 2023). To put the problem simply, Decker and Papagiannidis argue that “calls for XAI may underestimate the complexity necessary for AI models to be high functioning” (section of Decker and Papagiannidis in Budhwar et al. 2023, pp. 614–615). For example, the more complex the ML or AI models, the more difficult it is to obtain an appropriate, stable and non-deceptive approximation using post-hoc explanatory methods. Future researchers should pay close attention to innovative ideas in this area. A promising direction for further research is proposed by Deiseroth et al. (2023). They use a

modality-agnostic perturbation method (similar to PFI in this work) that manipulates the attentional mechanisms of transformer large language architectures to create relevance maps for the input with respect to the output prediction.

Second, missing ML transparency and comprehensibility would undermine the idea of the sociotechnical framework for fairness assessment to question the inner rationale of ML, respective generative AI models. Our understanding of alternative forms of validating, testing and approving models through an extensive evaluation of input-output tests and performance monitoring remains limited. In the data science community, some scholars and practitioners have been calling for this human-centered approach for years, for example when discussing the weaknesses of post-hoc explanatory methods in terms of their limitations (e.g., Kozyrkov 2018; Rudin 2019; Ghassemi et al. 2021; Vale et al. 2022). In these cases, it would be interesting to investigate whether the assessment of fairness, which depends on the transparency of the model (see Study 2), can be an assessment of fairness that is informed by extensive testing and validation. Of course, such a system would still be compatible with the six other requirements for high-risk decision-making in AI (see *Figure 4*). This approach offers interesting starting points for future research (e.g., what knowledge is required to test and validate opaque black-box AI models?; what framework and processes would lead to verifiable assurance of models?; how diversity, non-discrimination and fairness can be measured and improved through feedback loops during testing and validation?).

Third, the interface through which the results of generative AI systems are presented to the end-user differs significantly from the statistical measures and visualisation of the post-hoc explanation in this case study in study 3. Most generative AI systems (see, for example, ChatGPT) use natural language to enable chat interaction in an intuitive way. Nielsen suggests that AI is the first new user interface paradigm for 60 years (Nielsen 2023), moving beyond a command-based interaction (telling the computer what next step to take) to an intent-based outcome specification (telling the computer what outcome is wanted). In light of technology acceptance theory, new user interface paradigms should enable greater trust and acceptance of the model, due to perceived ease of use (Davis 1989). Thus, determinants for the adoption of the ML model in this case study – in terms of the theory of planned behaviour and ML characteristics – need to be adapted or extended (see Figure 19 in study 3). In this context, future research should scrutinise the effects of accessibility of different models as well as the acceptance and trust processes in interactive chat sessions when introducing AI systems in organisations from different stakeholders perspectives.

4 Conclusion

This dissertation provides an equally research-based perspective and practical insights into algorithmic HRM, highlighting the intriguing possibilities but also the alarming dangers of AI and ML technology in high-risk decision-making environments. The aim of this dissertation is to contribute to the current understanding of the literature on algorithmic HRM in order to shift the impact from the "dark" negative side to the "bright" positive side. It seems clear that (generative) AI will have an increasing impact on individuals, organisations and society as a whole in the coming years, and as a general-purpose technology it will impact all our lives. However, when looking deeper into the requirements for responsible use of AI, such as transparency and diversity, non-discrimination and fairness, it becomes clear that the negative impact that AI can bring requires attention from researchers, practitioners and society as a whole. In this sense, this dissertation is a minor advancement, but one that will hopefully encourage further developments towards the responsible use of AI in algorithmic HRM.

Publication bibliography

Adadi, Amina; Berrada, Mohammed (2018): Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI). In *IEEE Access* 6, pp. 52138–52160. DOI: 10.1109/ACCESS.2018.2870052.

Agrawal, Ajay; Gans, Joshua; Goldfarb, Avi (2018): Prediction machines. The simple economics of artificial intelligence. Boston, Mass.: Harvard Business Review Press.

AI HLEG EU (2019): High-Level Expert Group Artificial Intelligence (European Union) - Ethics guidelines for trustworthy AI. https://eskillsalliancecms.gov.mt/en/news/documents/2019/aidefinition.pdf.

Ajzen, Icek (1991): The theory of planned behavior. In *Organisational Behavior and Human Decision Processes* 50 (2), pp. 179–211. DOI: 10.1016/0749-5978(91)90020-t.

Ajzen, Icek (2002): Perceived Behavioral Control, Self-Efficacy, Locus of Control, and the Theory of Planned Behavior 1. In *Journal of Applied Social Psychology* 32 (4), pp. 665–683. DOI: 10.1111/j.1559-1816.2002.tb00236.x.

Ali, Sajid; Abuhmed, Tamer; El-Sappagh, Shaker; Muhammad, Khan; Alonso-Moral, Jose M.; Confalonieri, Roberto et al. (2023): Explainable Artificial Intelligence (XAI): What we know and what is left to attain Trustworthy Artificial Intelligence. In *Information Fusion* 99, p. 101805. DOI: 10.1016/j.inffus.2023.101805.

Alon-Barkat, Saar; Busuioc, Madalina (2023): Human–AI Interactions in Public Sector Decision Making: "Automation Bias" and "Selective Adherence" to Algorithmic Advice. In *Journal of Public Administration Research and Theory* 33 (1), pp. 153–169. DOI: 10.1093/jopart/muac007.

Angrave, David; Charlwood, Andy; Kirkpatrick, Ian; Lawrence, Mark; Stuart, Mark (2016): HR and analytics: why HR is set to fail the big data challenge. In *Human Resource Management Journal* 26 (1), pp. 1–11. DOI: 10.1111/1748-8583.12090.

Apley, Daniel W.; Zhu, Jingyu (2020): Visualizing the effects of predictor variables in black box supervised learning models. In *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 82 (4), pp. 1059–1086. DOI: 10.1111/rssb.12377.

Aral, Sinan; Brynjolfsson, Erik; Wu, Lynn (2012): Three-Way Complementarities: Performance Pay, Human Resource Analytics, and Information Technology. In *Management Science* 58 (5), pp. 913–931. DOI: 10.1287/mnsc.1110.1460.

Arrieta, Alejandro Barredo; Díaz-Rodríguez, Natalia; Del Ser, Javier; Bennetot, Adrien; Tabik, Siham; Barbado, Alberto et al. (2020): Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. In *Information Fusion* 58, pp. 82–115. DOI: 10.1016/j.inffus.2019.12.012.

Bader, Verena; Kaiser, Stephan (2019): Algorithmic decision-making? The user interface and its role for human involvement in decisions supported by artificial intelligence. In *Organisation* 26 (5), pp. 655–672. DOI: 10.1177/1350508419855714.

Bandura, A. (1977): Self-efficacy: toward a unifying theory of behavioral change. In *Psychological Review* 84 (2), pp. 191–215. DOI: 10.1037//0033-295x.84.2.191.

Basu, Shubhabrata; Majumdar, Bishakha; Mukherjee, Kajari; Munjal, Surender; Palaksha, Chandan (2023): Artificial Intelligence–HRM Interactions and Outcomes: A Systematic Review and Causal Configurational Explanation. In *Human Resource Management Review* 33 (1), p. 100893. DOI: 10.1016/j.hrmr.2022.100893.

Bauer, Kevin; Zahn, Moritz von; Hinz, Oliver (2023): Expl(AI)ned: The Impact of Explainable Artificial Intelligence on Users' Information Processing. In *Information Systems Research*, Article isre.2023.1199. DOI: 10.1287/isre.2023.1199.

Berger, Benedikt; Adam, Martin; Rühr, Alexander; Benlian, Alexander (2020): Watch Me Improve—Algorithm Aversion and Demonstrating the Ability to Learn. In *Business & information systems engineering*, pp. 1–14. DOI: 10.1007/s12599-020-00678-5.

Binns, Reuben; van Kleek, Max; Veale, Michael; Lyngs, Ulrik; Zhao, Jun; Shadbolt, Nigel (2018): 'It's Reducing a Human Being to a Percentage'; Perceptions of Justice in Algorithmic Decisions 16, pp. 1–14. DOI: 10.1145/3173574.3173951.

Böhmer, Nicole; Schinnenburg, Heike (2023): Critical exploration of AI-driven HRM to build up organisational capabilities. In *Employee Relations: The International Journal*. DOI: 10.1108/ER-04-2022-0202.

Bolukbasi, Tolga; Chang, Kai-Wei; Zou, James; Saligrama, Venkatesh; Kalai, Adam (2016): Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings. https://arxiv.org/pdf/1607.06520.pdf, updated on 2016.

Breiman, Leo (2001): Random Forests. In *Machine Learning* 45 (1), pp. 5–32. DOI: 10.1023/A:1010933404324.

Brown, Tom B.; Mann, Benjamin; Ryder, Nick; Subbiah, Melanie; Kaplan, Jared; Dhariwal, Prafulla et al. (2020): Language Models are Few-Shot Learners. https://arxiv.org/pdf/2005.14165.pdf.

Brynjolfsson, Erik; Li, Danielle; Raymond, Lindsey (2023): Generative AI at Work. In *Working Paper*. DOI: 10.3386/w31161.

Budhwar, Pawan; Malik, Ashish; Silva, M. T. Thedushika de; Thevisuthan, Praveena (2022): Artificial intelligence – challenges and opportunities for international HRM: a review and research agenda. In *The International Journal of Human Resource Management* 33 (6), pp. 1065–1097. DOI: 10.1080/09585192.2022.2035161.

Budhwar, Pawan S.; Chowdhury, Soumyadeb; Wood, Geoffrey T.; Aguinis, Herman; Bamber, Greg; Beltran, Jose R. et al. (2023): Human resource management in the age of generative artificial intelligence. Perspectives and research directions on ChatGPT. In *Human Resource Management Journal*. DOI: 10.1111/1748-8583.12524.

Burrell, Jenna (2016): How the machine 'thinks': Understanding opacity in machine learning algorithms. In *Big Data & Society* 3 (1). DOI: 10.1177/2053951715622512.

Busuioc, Madalina (2021): Accountable Artificial Intelligence: Holding Algorithms to Account. In *Public administration review* 81 (5), pp. 825–836. DOI: 10.1111/puar.13293.

Canhoto, Ana Isabel; Clear, Fintan (2020): Artificial intelligence and machine learning as business tools: A framework for diagnosing value destruction potential. In *Business Horizons* 63 (2), pp. 183–193. DOI: 10.1016/j.bushor.2019.11.003.

Castille, Christopher M.; Castille, Ann-Marie R. (2019): Disparate treatment and adverse impact in applied attrition modeling. In *Industrial and Organisational Psychology* 12 (3), pp. 310–313. DOI: 10.1017/iop.2019.53.

Charlwood, Andy; Guenole, Nigel (2022): Can HR adapt to the paradoxes of artificial intelligence? In *Human Resource Management Journal* 32 (4), Article 1748-8583.12433, pp. 729–742. DOI: 10.1111/1748-8583.12433.

Chatterjee, Sheshadri; Rana, Nripendra P.; Dwivedi, Yogesh K.; Baabdullah, Abdullah M. (2021): Understanding AI adoption in manufacturing and production firms using an integrated TAM-TOE model. In *Technological Forecasting and Social Change* 170, p. 120880. DOI: 10.1016/j.techfore.2021.120880.

Cheng, Maggie M.; Hackett, Rick D. (2021): A critical review of algorithms in HRM: Definition, theory, and practice. In *Human Resource Management Review* 31 (1), p. 100698. DOI: 10.1016/j.hrmr.2019.100698.

Choi, Jonathan H.; Hickman, Kristin E.; Monahan, Amy; Schwarcz, Daniel B. (2023): ChatGPT Goes to Law School. In *SSRN Electronic Journal*. DOI: 10.2139/ssrn.4335905.

Choudhury, Prithwiraj; Allen, Ryan T.; Endres, Michael G. (2021): Machine learning for pattern discovery in management research. In *Strategic Management Journal* 42 (1), pp. 30–57. DOI: 10.1002/smj.3215.

Chowdhury, Soumyadeb; Joel-Edgar, Sian; Dey, Prasanta Kumar; Bhattacharya, Sudeshna; Kharlamov, Alexander (2022): Embedding transparency in artificial intelligence machine learning models: managerial implications on predicting and explaining employee turnover. In *The International Journal of Human Resource Management* 34 (14), pp. 2732–2764. DOI: 10.1080/09585192.2022.2066981.

Chu, Mi-na (2023): Assessing the Benefits of ChatGPT for Business: An Empirical Study on Organisational Performance. In *IEEE Access*.

Civil Rights Act (1964): General Records of the United States Government, Record Group 11. https://www.archives.gov/milestone-documents/civil-rights-act.

Clark, Brent B.; Robert, Christopher; Hampton, Stephen A. (2016): The Technology Effect: How Perceptions of Technology Drive Excessive Optimism. In *Journal of Business and Psychology* 31 (1), pp. 87–102. DOI: 10.1007/s10869-015-9399-4.

Cohen, Jacob (1960): A Coefficient of Agreement for Nominal Scales. In *Educational and Psychological Measurement* 20 (1), pp. 37–46. DOI: 10.1177/001316446002000104.

Coolen, Patrick; van den Heuvel, Sjoerd; van de Voorde, Karina; Paauwe, Jaap (2023): Understanding the adoption and institutionalization of workforce analytics: A systematic literature review and research agenda. In *Human Resource Management Review*, p. 100985. DOI: 10.1016/j.hrmr.2023.100985.

Dastin, Jeffrey (2022): Amazon scraps secret AI recruiting tool that showed bias against women. In : Ethics of data and analytics: Auerbach Publications, pp. 296–299.

Daugherty, Paul R.; H. James Wilson; Rumman Chowdhury (2018): Using Artificial Intelligence to Promote Diversity. In *MIT Sloan Management Review*. https://sloanreview.mit.edu/article/using-artificial-intelligence-to-promote-diversity/.

Davenport, Thomas H. (2019): Is HR the Most Analytics-Driven Function? In *Harvard business review*. https://hbr.org/2019/04/is-hr-the-most-analytics-driven-function.

Davenport, Thomas H.; Bean, Randy (2024): Survey: GenAI Is Making Companies More Data Oriented. In *Harvard business review*. https://hbr.org/2024/01/survey-genai-is-making-companies-more-data-oriented, checked on 1/21/2024.

Davenport, Thomas H.; Harris, Jeanne; Shapiro, Jeremy (2010): Competing on talent analytics. In *Harvard business review* 88 (10), 52-8, 150. https://www.researchgate.net/publication/47369355_Competing_on_talent_analytics.

Davis, Fred D. (1989): Perceived Usefulness, Perceived Ease of Use, and User Acceptance of Information Technology. In *MIS Quarterly* 13 (3), p. 319. DOI: 10.2307/249008.

Deiseroth, Björn; Deb, Mayukh; Weinbach, Samuel; Brack, Manuel; Schramowski, Patrick; Kersting, Kristian (2023): AtMan: Understanding Transformer Predictions Through Memory Efficient Attention Manipulation.

Desouza, Kevin C.; Dawson, Gregory S.; Chenok, Daniel (2020): Designing, developing, and deploying artificial intelligence systems: Lessons from and for the public sector. In *Business Horizons* 63 (2), pp. 205–213. DOI: 10.1016/j.bushor.2019.11.004.

Díaz-Rodríguez, Natalia; Del Ser, Javier; Coeckelbergh, Mark; López de Prado, Marcos; Herrera-Viedma, Enrique; Herrera, Francisco (2023): Connecting the dots in trustworthy Artificial

Intelligence: From AI principles, ethics, and key requirements to responsible AI systems and regulation. In *Information Fusion*, p. 101896. DOI: 10.1016/j.inffus.2023.101896.

Dietvorst, B. J.; Simmons, J. P.; Massey, C. (2018): Overcoming algorithm aversion: People will use imperfect algorithms if they can (even slightly) modify them. In *Management Science* 64 (3), pp. 1155–1170. DOI: 10.1287/mnsc.2016.2643.

Dodge, Jonathan; Liao, Q. Vera; Zhang, Yunfeng; Bellamy, Rachel K. E.; Dugan, Casey (2019): Explaining models. In Wai-Tat Fu, Shimei Pan, Oliver Brdiczka, Polo Chau, Gaelle Calvary (Eds.): Proceedings of the 24th International Conference on Intelligent User Interfaces. IUI '19: 24th International Conference on Intelligent User Interfaces. Marina del Ray California. New York, NY, USA: ACM, pp. 275–285.

Dolata, Mateusz; Feuerriegel, Stefan; Schwabe, Gerhard (2022): A sociotechnical view of algorithmic fairness. In *Information Systems Journal* 32 (4), pp. 754–818. DOI: 10.1111/isj.12370.

Downes, Patrick E.; Harris, T. Brad; Allen, David G. (2023): Getting from valid to useful: End user modifiability and human capital analytics implementation in selection. In *Human Resource Management*, Article hrm.22179. DOI: 10.1002/hrm.22179.

Dwivedi, Yogesh K.; Hughes, Laurie; Ismagilova, Elvira; Aarts, Gert; Coombs, Crispin; Crick, Tom et al. (2021): Artificial Intelligence (AI): Multidisciplinary perspectives on emerging challenges, opportunities, and agenda for research, practice and policy. In *International Journal of Information Management* 57, p. 101994. DOI: 10.1016/j.ijinfomgt.2019.08.002.

Dwivedi, Yogesh K.; Kshetri, Nir; Hughes, Laurie; Slade, Emma Louise; Jeyaraj, Anand; Kar, Arpan Kumar et al. (2023): "So what if ChatGPT wrote it?" Multidisciplinary perspectives on opportunities, challenges and implications of generative conversational AI for research, practice and policy. In *International Journal of Information Management* 71, p. 102642. DOI: 10.1016/j.ijinfomgt.2023.102642.

Edwards, Lilian (2022): The EU AI Act: a summary of its significance and scope. Newcastle University. https://www.adalovelaceinstitute.org/wp-content/uploads/2022/04/Expert-explainer-The-EU-AI-Act-11-April-2022.pdf, checked on 11/28/2023.

Edwards, Martin R.; Charlwood, Andy; Guenole, Nigel; Marler, Janet (2022): HR analytics: An emerging field finding its place in the world alongside simmering ethical challenges. In *Human Resource Management Journal*, Article 1748-8583.12435. DOI: 10.1111/1748-8583.12435.

Einola, Katja; Khoreva, Violetta (2023): Best friend or broken tool? Exploring the co-existence of humans and artificial intelligence in the workplace ecosystem. In *Human Resource Management* 62 (1), pp. 117–135. DOI: 10.1002/hrm.22147.

Ellmer, M.; Reichel, A. (2021): Staying close to business: the role of epistemic alignment in rendering HR analytics outputs relevant to decision-makers. In *International Journal of Human Resource Management* 32 (12), pp. 2622–2642. DOI: 10.1080/09585192.2021.1886148.

Epstein, Ziv; Hertzmann, Aaron; Akten, Memo; Farid, Hany; Fjeld, Jessica; Frank, Morgan R. et al. (2023): Art and the science of generative AI. In *Science (New York, N.Y.)* 380 (6650), pp. 1110–1111. DOI: 10.1126/science.adh4451.

Erel, Isil; Stern, Léa H.; Tan, Chenhao; Weisbach, Michael S. (2021): Selecting Directors Using Machine Learning. In *The Review of Financial Studies* 34 (7), pp. 3226–3264. DOI: 10.1093/rfs/hhab050.

EU Council (2023): Artificial intelligence act: Council and Parliament strike a deal on the first rules for AI in the world. https://www.consilium.europa.eu/en/press/press-releases/2023/12/09/artificial-intelligence-act-council-and-parliament-strike-a-deal-on-the-first-worldwide-rules-for-ai/, updated on 12/15/2023, checked on 12/15/2023.

European Parliament (2022): Parliamentary question: The Dutch childcare benefit scandal, institutional racism and algorithms | O-000028/2022.

https://www.europarl.europa.eu/doceo/document/O-9-2022-000028_EN.html, updated on 11/27/2023, checked on 11/27/2023.

European Parliament (2024): EU AI Act: first regulation on artificial intelligence. https://www.europarl.europa.eu/news/en/headlines/society/20230601STO93804/eu-ai-act-first-regulation-on-artificial-intelligence, checked on 11/27/2024.

Fauzi, Fauzi; Tuhuteru, Laros; Sampe, Ferdinandus; Ausat, Abu Muna Almaududi; Hatta, Heliza Rahmania (2023): Analysing the Role of ChatGPT in Improving Student Productivity in Higher Education. In *Journal on Education* 5 (4), pp. 14886–14891. DOI: 10.31004/joe.v5i4.2563.

Fest, Isabelle; Schäfer, Mirko; van Dijck, José; Meijer, Albert (2023): Understanding Data Professionals in the Police: A Qualitative Study of System-Level Bureaucrats. In *Public Management Review*, pp. 1–21. DOI: 10.1080/14719037.2023.2222734.

Fishbein, Martin; Ajzen, Icek (2010): Predicting and changing behavior: The reasoned action approach. New York, NY, US: Psychology Press (Predicting and changing behavior: The reasoned action approach).

Flyvbjerg, Bent (2006): Five Misunderstandings About Case-Study Research. In *Qualitative Inquiry* 12 (2), pp. 219–245. DOI: 10.1177/1077800405284363.

Friedman, Nicola; Ormiston, Jarrod (2022): Blockchain as a sustainability-oriented innovation?: Opportunities for and resistance to Blockchain technology as a driver of sustainability in global food supply chains. In *Technological Forecasting and Social Change* 175, p. 121403. DOI: 10.1016/j.techfore.2021.121403.

Gal, Uri; Jensen, Tina Blegind; Stein, Mari-Klara (2020): Breaking the vicious cycle of algorithmic management: A virtue ethics approach to people analytics. In *Information and Organisation* 30 (2), p. 100301. DOI: 10.1016/j.infoandorg.2020.100301.

Gartner (2024): Hype Cycle for Artificial Intelligence 2024, https://www.gartner.com/en/articles/hype-cycle-for-artificial-intelligence, updated on 11/15/2024, checked on 09/01/2025.

GDPR (2016): General Data Protection Regulation of the European Union, Article 5(1)(a). Regulation (EU) 2016/679 of the European Parliament and of the Council on the protection of natural persons with regard to the processing of personal data and on the free movement of such data. https://eur-lex.europa.eu/eli/reg/2016/679/oj, updated on 12/19/2023, checked on 12/19/2023.

German Network Artificial Intelligence in Labour and Social Administration (2022): A value guideline for the use of AI in labor and social administration. https://www.bmas.de/SharedDocs/Downloads/DE/Publikationen/a862-01-leitlinien-ki-einsatz-behoerdliche-praxis-arbeits-sozialverwaltung.pdf?__blob=publicationFile&v=2.

Ghassemi, Marzyeh; Oakden-Rayner, Luke; Beam, Andrew L. (2021): The false hope of current approaches to explainable artificial intelligence in health care. In *The Lancet. Digital health* 3 (11), 745-750. DOI: 10.1016/S2589-7500(21)00208-9.

Gieter, Sara de; Cooman, Rein de; Hofmans, Joeri; Pepermans, Roland; Jegers, Marc (2012): Pay-Level Satisfaction and Psychological Reward Satisfaction as Mediators of the Organisational Justice-Turnover Intention Relationship. In *International Studies of Management & Organisation* 42 (1), pp. 50–67. DOI: 10.2753/IMO0020-8825420103.

Gilson, Aidan; Safranek, Conrad W.; Huang, Thomas; Socrates, Vimig; Chi, Ling; Taylor, Richard Andrew; Chartash, David (2023): How does ChatGPT perform on the United States medical licensing examination? The implications of large language models for medical education and knowledge assessment. In *JMIR Medical Education* 9 (1), e45312.

Gioia, Dennis A.; Corley, Kevin G.; Hamilton, Aimee L. (2013): Seeking Qualitative Rigor in Inductive Research. In *Organisational Research Methods* 16 (1), pp. 15–31. DOI: 10.1177/1094428112452151.

Glikson, Ella; Woolley, Anita Williams (2020): Human Trust in Artificial Intelligence: Review of Empirical Research. In *Academy of Management Annals* 14 (2), pp. 627–660. DOI: 10.5465/annals.2018.0057.

Golgeci, Ismail; Ritala, Paavo; Arslan, Ahmad; McKenna, Brad; Ali, Imran (2025): Confronting and alleviating AI resistance in the workplace: An integrative review and a process. In *Human Resource Management Review* 35 (2), p. 101075. DOI: 10.1016/j.hrmr.2024.101075

Goodman, Bryce; Flaxman, Seth (2017): European Union regulations on algorithmic decision-making and a "right to explanation". In *AI Magazine* 38 (3), pp. 50–57.

Gray, Alastair M.; Phillips, V. L. (1994): Turnover, age and length of service: a comparison of nurses and other staff in the National Health Service. In *Journal of advanced nursing* 19 (4), pp. 819–827. DOI: 10.1111/j.1365-2648.1994.tb01155.x.

Gregor, Shirley; Hevner, Alan R. (2013): Positioning and presenting design science research for maximum impact. In *Management information systems*.

Griffin, Andrew (2023): ChatGPT creators try to use artificial intelligence to explain itself – and come across major problems. In *The Independent*, 5/12/2023. https://www.independent.co.uk/tech/chatgpt-website-openai-artificial-intelligence-b2337503.html, checked on 1/13/2024.

Grissom, Jason A.; Viano, Samantha L.; Selin, Jennifer L. (2016): Understanding Employee Turnover in the Public Sector: Insights from Research on Teacher Mobility. In *Public Administration Review* 76 (2), pp. 241–251. DOI: 10.1111/puar.12435.

Gunning, David; Aha, David W. (2019): DARPA's Explainable Artificial Intelligence Program. In *AI Magazine* 40 (2), pp. 44–58. DOI: 10.1609/aimag.v40i2.2850.

Ha, Taehyun; Kim, Sangyeon (2023): Improving Trust in AI with Mitigating Confirmation Bias: Effects of Explanation Type and Debiasing Strategy for Decision-Making with Explainable AI. In *International Journal of Human–Computer Interaction*, pp. 1–12. DOI: 10.1080/10447318.2023.2285640.

Haque, A. BahalulK.M.; Islam, A.K.M. Najmul; Mikalef, Patrick (2023): Explainable Artificial Intelligence (XAI) from a user perspective: A synthesis of prior literature and problematizing avenues for future research. In *Technological Forecasting and Social Change* 186, p. 122120. DOI: 10.1016/j.techfore.2022.122120.

Hickman, L.; Saef, R.; Ng, V.; Woo, S. E.; Tay, L.; Bosch, N. (2021): Developing and evaluating language-based machine learning algorithms for inferring applicant personality in video interviews. In *Human Resource Management Journal*. DOI: 10.1111/1748-8583.12356.

Holstein, Kenneth; Wortman Vaughan, Jennifer; Daumé, Hal; Dudik, Miro; Wallach, Hanna (2019): Improving Fairness in Machine Learning Systems. In Stephen Brewster, Geraldine Fitzpatrick, Anna Cox, Vassilis Kostakos (Eds.): Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems. CHI '19: CHI Conference on Human Factors in Computing Systems. Glasgow Scotland Uk. New York, NY, USA: ACM, pp. 1–16.

Holtom, Brooks C.; Mitchell, Terence R.; Lee, Thomas W.; Eberly, Marion B. (2008): 5 Turnover and Retention Research: A Glance at the Past, a Closer Review of the Present, and a Venture into the Future. In *The Academy of Management Annals* 2 (1), pp. 231–274. DOI: 10.1080/19416520802211552.

Holzwarth, Sebastian; Gunnesch-Luca, George; Soucek, Roman; Moser, Klaus (2021): How Communication in Organisations Is Related to Foci of Commitment and Turnover Intentions. In *Journal of Personnel Psychology* 20 (1), pp. 27–38. DOI: 10.1027/1866-5888/a000261.

Hunkenschroer, Anna Lena; Luetge, Christoph (2022): Ethics of AI-Enabled Recruiting and Selection: A Review and Research Agenda. In *Journal of Business Ethics*. DOI: 10.1007/s10551-022-05049-6.

Jo, Hyeon; Park, Do-Hyung (2023): AI in the Workplace: Examining the Effects of ChatGPT on Information Support and Knowledge Acquisition. In *International Journal of Human–Computer Interaction*, pp. 1–16. DOI: 10.1080/10447318.2023.2278283.

Kaddour, Jean; Lynch, Aengus; Liu, Qi; Kusner, Matt J.; Silva, Ricardo (2022): Causal Machine Learning: A Survey and Open Problems.

Kelan, Elisabeth K. (2023): Algorithmic inclusion: Shaping the predictive algorithms of artificial intelligence in hiring. In *Human Resource Management Journal*, Article 1748-8583.12511. DOI: 10.1111/1748-8583.12511.

Kellogg, Katherine C.; Valentine, Melissa A.; Christin, Angéle (2020): Algorithms at Work: The New Contested Terrain of Control. In *Academy of Management Annals* 14 (1), pp. 366–410. DOI: 10.5465/annals.2018.0174.

Kim, Doha; Song, Yeosol; Kim, Songyie; Lee, Sewang; Wu, Yanqin; Shin, Jungwoo; Lee, Daeho (2023): How should the results of artificial intelligence be explained to users? - Research on consumer preferences in user-centered explainable artificial intelligence. In *Technological Forecasting and Social Change* 188, p. 122343. DOI: 10.1016/j.techfore.2023.122343.

Kim, S.; Wang, Y.; Boon, C. (2021): Sixty years of research on technology and human resource management: Looking back and looking forward. In *Human Resource Management* 60 (1), pp. 229–247. DOI: 10.1002/hrm.22049.

King, Kylie Goodell (2016): Data Analytics in Human Resources. In *Human Resource Development Review* 15 (4), pp. 487–495. DOI: 10.1177/1534484316675818.

Köchling, Alina; Wehner, Marius Claus (2020): Discriminated by an algorithm: a systematic review of discrimination and fairness by algorithmic decision-making in the context of HR recruitment and HR development. In *Business Research* 13 (3), pp. 795–848. DOI: 10.1007/s40685-020-00134-w.

Kordzadeh, Nima; Ghasemaghaei, Maryam (2022): Algorithmic bias: review, synthesis, and future research directions. In *European Journal of Information Systems* 31 (3), pp. 388–409. DOI: 10.1080/0960085X.2021.1927212.

Kozyrkov, Cassie (2018): Explainable AI won't deliver. Here's why. https://kozyrkov.medium.com/explainable-ai-wont-deliver-here-s-why-6738f54216be, checked on 1/12/2024.

Krishna, Satyapriya; Ma, Jiaqi; Lakkaraju, Himabindu (2023): Towards Bridging the Gaps between the Right to Explanation and the Right to be Forgotten. http://arxiv.org/pdf/2302.04288v2.

Kuhn, Max (2019): Building predictive models in R using the caret package. https://topepo.github.io/caret/index.html, updated on 3/27/2019, checked on 12/30/2021.

Lambrecht, Anja; Tucker, Catherine (2019): Algorithmic Bias? An Empirical Study of Apparent Gender-Based Discrimination in the Display of STEM Career Ads. In *Management Science* 65 (7), pp. 2966–2981. DOI: 10.1287/mnsc.2018.3093.

Lamers, Laura; Meijerink, Jeroen; Rettagliata, Giorgio (2024): Blinded by "algo economicus": Reflecting on the assumptions of algorithmic management research to move forward. In *Human Resource Management*, Article hrm.22204. DOI: 10.1002/hrm.22204.

Landis, J. Richard; Koch, Gary G. (1977): An Application of Hierarchical Kappa-type Statistics in the Assessment of Majority Agreement among Multiple Observers. In *Biometrics* 33 (2), p. 363. DOI: 10.2307/2529786.

Langer, Markus; Hunsicker, Tim; Feldkamp, Tina; König, Cornelius J.; Grgić-Hlača, Nina (2022): "Look! It's a Computer Program! It's an Algorithm! It's AI!": Does Terminology Affect Human Perceptions and Evaluations of Algorithmic Decision-Making Systems? In Simone Barbosa, Cliff Lampe, Caroline Appert, David A. Shamma, Steven Drucker, Julie Williamson, Koji Yatani (Eds.): CHI Conference on Human Factors in Computing Systems. CHI '22: CHI Conference on Human

Factors in Computing Systems. New Orleans LA USA, 29 04 2022 05 05 2022. New York, NY, USA: ACM, pp. 1–28.

Langer, Markus; König, Cornelius J. (2023): Introducing a multi-stakeholder perspective on opacity, transparency and strategies to reduce opacity in algorithm-based human resource management. In *Human Resource Management Review* 33 (1), p. 100881. DOI: 10.1016/j.hrmr.2021.100881.

Langer, Markus; König, Cornelius J.; Back, Caroline; Hemsing, Victoria (2023): Trust in Artificial Intelligence: Comparing Trust Processes Between Human and Automated Trustees in Light of Unfair Bias. In *Journal of Business and Psychology* 38 (3), pp. 493–508. DOI: 10.1007/s10869-022-09829-9.

Lawrance, N.; Petrides, G.; Guerry, M.-A. (2021): Predicting employee absenteeism for cost effective interventions. In *Decision Support Systems* 147. DOI: 10.1016/j.dss.2021.113539.

Leavitt, Keith; Schabram, Kira; Hariharan, Prashanth; Barnes, Christopher M. (2021): Ghost in the Machine: On Organisational Theory in the Age of Machine Learning. In *Academy of Management Review* 46 (4), pp. 750–777. DOI: 10.5465/amr.2019.0247.

Lee; Floridi, Luciano; Singh, Jatinder (2021): Formalising trade-offs beyond algorithmic fairness: lessons from ethical philosophy and welfare economics. In *AI and Ethics* 1 (4), pp. 529–544. DOI: 10.1007/s43681-021-00067-y.

Lee, ChangHyun; Cha, KyungJin (2023): FAT-CAT—Explainability and augmentation for an AI system: A case study on AI recruitment-system adoption. In *International Journal of Human-Computer Studies* 171, p. 102976. DOI: 10.1016/j.ijhcs.2022.102976.

Lee, Min Kyung (2018): Understanding perception of algorithmic decisions: Fairness, trust, and emotion in response to algorithmic management. In *Big Data & Society* 5 (1), 205395171875668. DOI: 10.1177/2053951718756684.

Leicht-Deobald, Ulrich; Busch, Thorsten; Schank, Christoph; Weibel, Antoinette; Schafheitle, Simon; Wildhaber, Isabelle; Kasper, Gabriel (2019): The Challenges of Algorithm-Based HR Decision-Making for Personal Integrity. In *Journal of business ethics : JBE* 160 (2), pp. 377–392. DOI: 10.1007/s10551-019-04204-w.

Leitner-Hanetseder, Susanne; Lehner, Othmar M.; Eisl, Christoph; Forstenlechner, Carina (2021): A profession in transition: actors, tasks and roles in AI-based accounting. In *Journal of Applied Accounting Research* 22 (3), pp. 539–556. DOI: 10.1108/JAAR-10-2020-0201.

Lin, Li; Bai, Yuntao; Mo, Changwei; Liu, Dong; Li, Xiyuan (2021): Does pay raise decrease temporary agency workers' voluntary turnover over time in China? Understanding the moderating role of demographics. In *The International Journal of Human Resource Management* 32 (7), pp. 1537–1565. DOI: 10.1080/09585192.2018.1539861.

Lum, Lillie; Kervin, John; Clark, Kathleen; Reid, Frank; Sirola, Wendy (1998): Explaining nursing turnover intent: job satisfaction, pay satisfaction, or organisational commitment? In *Journal of Organisational Behavior* 19 (3), pp. 305–320. DOI: 10.1002/(SICI)1099-1379(199805)19:3<305::AID-JOB843>3.0.CO;2-N.

Lundberg, Scott; Lee, Su-In (2017): A Unified Approach to Interpreting Model Predictions. In *Proceedings of the 31st international conference on neural information processing system*. http://arxiv.org/pdf/1705.07874v2.

Mahmud, Hasan; Islam, A.K.M. Najmul; Ahmed, Syed Ishtiaque; Smolander, Kari (2022): What influences algorithmic decision-making? A systematic literature review on algorithm aversion. In *Technological Forecasting and Social Change* 175, p. 121390. DOI: 10.1016/j.techfore.2021.121390.

Malik, Ashish; Budhwar, Pawan; Kazmi, Bahar Ali (2023): Artificial intelligence (AI)-assisted HRM: Towards an extended strategic framework. In *Human Resource Management Review* 33 (1), p. 100940. DOI: 10.1016/j.hrmr.2022.100940.

Margherita, A. (2021): Human resources analytics: A systematization of research topics and directions for future research. In *Human Resource Management Review* 32 (2), Article 100795. DOI: 10.1016/j.hrmr.2020.100795.

Marler, Janet H.; Boudreau, John W. (2017): An evidence-based review of HR Analytics. In *International Journal of Human Resource Management* 28 (1), pp. 3–26. DOI: 10.1080/09585192.2016.1244699.

McCartney, Steven; Fu, Na (2022): Bridging the gap: why, how and when HR analytics can impact organisational performance. In *Management Decision* 60 (13), pp. 25–47. DOI: 10.1108/MD-12-2020-1581.

Meddeb, Eya; Bowers, Christopher; Nichol, Lynn (2022): Comparing Machine Learning Correlations to Domain Experts' Causal Knowledge: Employee Turnover Use Case. In Andreas Holzinger (Ed.): Machine Learning and Knowledge Extraction: 6th IFIP TC 5, TC 12, WG 8.4, WG 8.9, WG 12.9 International Cross-Domain Conference, CD-MAKE 2022, Vienna, Austria, August 23-26, 2022, Proceedings. International Cross-Domain Conference for Machine Learning and Knowledge Extraction: Springer, pp. 343–361. https://link.springer.com/chapter/10.1007/978-3-031-14463-9_22.

Mehrabi, Ninareh; Morstatter, Fred; Saxena, Nripsuta; Lerman, Kristina; Galstyan, Aram (2019): A Survey on Bias and Fairness in Machine Learning. http://arxiv.org/pdf/1908.09635v3.

Meijerink, Jeroen; Bondarouk, Tanya (2021): The duality of algorithmic management: Toward a research agenda on HRM algorithms, autonomy and value creation. In *Human Resource Management Review*, p. 100876. DOI: 10.1016/j.hrmr.2021.100876.

Meijerink, Jeroen; Boons, Mark; Keegan, Anne; Marler, Janet (2021): Algorithmic human resource management: Synthesizing developments and cross-disciplinary insights on digital HRM. In *The International Journal of Human Resource Management* 32 (12), pp. 2545–2562. DOI: 10.1080/09585192.2021.1925326.

Meijerink, Jeroen; Keegan, Anne (2019): Conceptualizing human resource management in the gig economy. In *Journal of Managerial Psychology* 34 (4), pp. 214–232. DOI: 10.1108/JMP-07-2018-0277.

Merton, Robert K.; Kendall, Patricia L. (1946): The Focused Interview. In *American Journal of Sociology* 51 (6), pp. 541–557. DOI: 10.1086/219886.

Meyer, David (2018): Amazon Reportedly Killed an AI Recruitment System Because It Couldn't Stop the Tool from Discriminating Against Women. https://fortune.com/2018/10/10/amazon-ai-recruitment-bias-women-sexist/.

Mikalef, Patrick; Conboy, Kieran; Lundström, Jenny Eriksson; Popovič, Aleš (2022): Thinking responsibly about responsible AI and 'the dark side' of AI. In *European Journal of Information Systems* 31 (3), pp. 257–268. DOI: 10.1080/0960085X.2022.2026621.

Miles, Matthew B.; Huberman, A. Michael (1994): Qualitative data analysis. An expanded sourcebook. 2. ed [Nachdr.]. Thousand Oaks, Calif.: Sage.

Mökander, Jakob; Axente, Maria; Casolari, Federico; Floridi, Luciano (2022): Conformity Assessments and Post-market Monitoring: A Guide to the Role of Auditing in the Proposed European AI Regulation. In *Minds and machines* 32 (2), pp. 241–268. DOI: 10.1007/s11023-021-09577-4.

Molnar, Christoph (2022): Interpretable machine learning. A guide for making Black Box Models interpretable. 2. Edition. Morisville, North Carolina: Lulu.

Molnar, Christoph; Casalicchio, Giuseppe; Bischl, Bernd (2018): iml: An R package for interpretable machine learning. In *Journal of Open Source Software* 3 (26), p. 786. https://doi.org/10.21105/joss.00786.

Moon, Ken; Loyalka, Prashant; Bergemann, Patrick; Cohen, Joshua (2022): The Hidden Cost of Worker Turnover: Attributing Product Reliability to the Turnover of Factory Workers. In *Management Science* 68 (5), pp. 3755–3767. DOI: 10.1287/mnsc.2022.4311.

Morgenstern, Oskar; Rubinstein, Ariel; Neumann, John von (2007): Theory of Games and Economic Behavior (60th Anniversary Commemorative Edition). With assistance of Harold William Kuhn. Princeton, N.J: Princeton University Press.

Mula, Claire; Zybura, Nora; Hipp, Thomas (2024): From digitalized start-up to scale-up: Opening the black box of scaling in digitalized firms towards a scaling process framework. In *Technological Forecasting and Social Change* 202, p. 123275. DOI: 10.1016/j.techfore.2024.123275.

Network AI in labour and social administration (2022): Self-imposed guideflines for the use of AI in the official practice of labor and social administration. https://bghm-magazin.de/fileadmin/user_upload/BGHM/Presseportal/Fachartikel2022/Selbstverpflichtende-Leitlinien-Kuenstliche-Intelligenz.pdf.

Neumann, Oliver; Guirguis, Katharina; Steiner, Reto (2022): Exploring artificial intelligence adoption in public organisations: a comparative case study. In *Public Management Review*, pp. 1–28. DOI: 10.1080/14719037.2022.2048685.

Newman, D. T.; Fast, N. J.; Harmon, D. J. (2020): When eliminating bias isn't fair: Algorithmic reductionism and procedural justice in human resource decisions. In *Organisational Behavior and Human Decision Processes* 160, pp. 149–167.

Nielsen, Jakob (2023): AI Is First New UI Paradigm in 60 Years. https://www.uxtigers.com/post/ai-new-ui-paradigm, updated on 10/24/2023, checked on 1/12/2024.

Noy, Shakked; Zhang, Whitney (2023): Experimental evidence on the productivity effects of generative artificial intelligence. In *Science (New York, N.Y.)* 381 (6654), pp. 187–192. DOI: 10.1126/science.adh2586.

Omrani, Nessrine; Rivieccio, Giorgia; Fiore, Ugo; Schiavone, Francesco; Agreda, Sergio Garcia (2022): To trust or not to trust? An assessment of trust in AI-based systems: Concerns, ethics and contexts. In *Technological Forecasting and Social Change* 181, p. 121763. DOI: 10.1016/j.techfore.2022.121763.

O'Neil, Cathy (2016): Weapons of math destruction. How big data increases inequality and threatens democracy. First edition. New York: Crown.

Park, Hyanghee; Ahn, Daehwan; Hosanagar, Kartik; Lee, Joonhwan (2021): Human-AI Interaction in Human Resource Management: Understanding Why Employees Resist Algorithmic Evaluation at Workplaces and How to Mitigate Burdens. In Yoshifumi Kitamura, Aaron Quigley, Katherine Isbister, Takeo Igarashi, Pernille Bjørn, Steven Drucker (Eds.): Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems. CHI '21: CHI Conference on Human Factors in Computing Systems. Yokohama Japan, 08 05 2021 13 05 2021. New York, NY, USA: ACM, pp. 1–15.

Pasquale, Frank. (2015): The black box society: The secret algorithms that control money and information: Harvard University Press. https://www.degruyter.com/document/doi/10.4159/harvard.9780674736061.c8/html.

Peng, Sida; Kalliamvakou, Eirini; Cihon, Peter; Demirer, Mert (2023): The Impact of AI on Developer Productivity: Evidence from GitHub Copilot. http://arxiv.org/pdf/2302.06590v1.

Pickard, Matthew D.; Cokins, Gary (2015): From Bean Counters to Bean Growers: Accountants as Data Analysts—A Customer Profitability Example. In *Journal of Information Systems* 29 (3), pp. 151–164. DOI: 10.2308/isys-51180.

PricewaterhouseCoopers (2024): Thriving in an age of continuous reinvention. https://www.pwc.com/gx/en/issues/c-suite-insights/ceo-survey.html, updated on 1/21/2024, checked on 1/21/2024.

Prikshat, V.; Malik, A.; Budhwar, P. (2023a): AI-augmented HRM: Antecedents, assimilation and multilevel consequences. In *Human Resource Management Review* (33), p. 100860. DOI: 10.1016/j.hrmr.2021.100860.

Prikshat, Verma; Islam, Mohammad; Patel, Parth; Malik, Ashish; Budhwar, Pawan; Gupta, Suraksha (2023b): AI-Augmented HRM: Literature review and a proposed multilevel framework for future research. In *Technological Forecasting and Social Change* 193, p. 122645. DOI: 10.1016/j.techfore.2023.122645.

Provost, Foster; Fawcett, Tom (2013): Data Science for Business. 1st ed. Beijing, Cambridge, Farnham, Köln, Sebastopol, Tokyo: O'Really.

Pumplun, Luisa; Tauchert, Christoph; Heidt, Margareta (2019): A new organisational chassis for artificial intelligence - exploring organisational readiness factors. In : ECIS 2019 proceedings. Stockholm, Uppsala: Association for Information Systems. https://aisel.aisnet.org/ecis2019_rp/106.

Putka, Dan J.; Beatty, Adam S.; Reeder, Matthew C. (2018): Modern Prediction Methods: New Perspectives on a Common Problem. In *Organisational Research Methods* 21 (3), pp. 689–732. DOI: 10.1177/1094428117697041.

R Core Team, R. (2013): R: A language and environment for statistical computing.

Raisch, Sebastian; Krakowski, Sebastian (2021): Artificial Intelligence and Management: The Automation–Augmentation Paradox. In *Academy of Management Review* 46 (1), pp. 192–210. DOI: 10.5465/amr.2018.0072.

Rasmussen, Thomas; Ulrich, Dave (2015): Learning from practice: How HR analytics avoids being a management fad. In *Organisational Dynamics* 44 (3), pp. 236–242. DOI: 10.1016/j.orgdyn.2015.05.008.

Reich, Taly; Kaju, Alex; Maglio, Sam J. (2023): How to overcome algorithm aversion: Learning from mistakes. In *Journal of Consumer Psychology* 33 (2), pp. 285–302. DOI: 10.1002/jcpy.1313.

Remneland Wikhamn, Björn; Styhre, Alexander; Wikhamn, Wajda (2023): HRM work and open innovation: evidence from a case study. In *The International Journal of Human Resource Management* 34 (10), pp. 1940–1972. DOI: 10.1080/09585192.2022.2054285.

Ribeiro, Marco Tulio; Singh, Sameer; Guestrin, Carlos (2016): " Why should i trust you?" Explaining the predictions of any classifier. In : Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining, pp. 1135–1144.

Riveiro, Maria; Thill, Serge (2021): “That's (not) the output I expected!” On the role of end user expectations in creating explanations of AI systems. In *Artificial Intelligence* 298, p. 103507. DOI: 10.1016/j.artint.2021.103507.

Rogers, Everett M. (2003): Diffusion of innovations. Fifth edition, Free Press trade paperback edition. New York, London, Toronto, Sydney: Free Press (Social science). http://www.loc.gov/catdir/bios/simon052/2003049022.html.

Rombaut, Evy; Guerry, Marie-Anne (2018): Predicting voluntary turnover through human resources database analysis. In *Management Research Review* 41 (1), pp. 96–112. DOI: 10.1108/MRR-04-2017-0098.

Rombaut, Evy; Guerry, Marie-Anne (2021): Determinants of voluntary turnover: A data-driven analysis for blue and white collar workers. In *Work (Reading, Mass.)* 69 (3), pp. 1083–1101. DOI: 10.3233/WOR-213538.

Rottman, Caleb; Gardner, Cari; Liff, Joshua; Mondragon, Nathan; Zuloaga, Lindsey (2023): New strategies for addressing the diversity-validity dilemma with big data. In *The Journal of Applied Psychology* 108 (9), pp. 1425–1444. DOI: 10.1037/apl0001084.

Rubenstein, Alex L.; Eberly, Marion B.; Lee, Thomas W.; Mitchell, Terence R. (2018): Surveying the forest: A meta-analysis, moderator investigation, and future-oriented discussion of the antecedents of voluntary employee turnover. In *Personnel Psychology* 71 (1), pp. 23–65. DOI: 10.1111/peps.12226.

Rudin, Cynthia (2019): Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead. In *Nature Machine Intelligence* 1 (5), pp. 206–215. DOI: 10.1038/s42256-019-0048-x.

Rudin, Cynthia; Wang, Caroline; Coker, Beau (2020): The age of secrecy and unfairness in recidivism prediction. In *Harvard Data Science Review* 2 (1), p. 1.

Russell, Craig J.; Sell, Mary V. (2012): A closer look at decisions to quit. In *Organisational Behavior and Human Decision Processes* 117 (1), pp. 125–137. DOI: 10.1016/j.obhdp.2011.09.002.

Samuel, Arthur L. (1959): Some studies in machine learning using the game of checkers. In *IBM Journal of research and development* 3 (3), pp. 210–229.

Sarker, Suprateek; Chatterjee, Sutirtha; Xiao, Xiao; Elbanna, Amany (2019): The Sociotechnical Axis of Cohesion for the IS Discipline: Its Historical Legacy and its Continued Relevance. In *MIS Quarterly* 43 (3), pp. 695–719. DOI: 10.25300/MISQ/2019/13747.

Schmidt, Philipp; Biessmann, Felix; Teubner, Timm (2020): Transparency and trust in artificial intelligence systems. In *Journal of Decision Systems* 29 (4), pp. 260–278. DOI: 10.1080/12460125.2020.1819094.

Schuessler, Elke S.; Lohmeyer, Nora; Ashwin, Sarah (2023): "We Can't Compete on Human Rights": Creating Market-Protected Spaces to Institutionalize the Emerging Logic of Responsible Management. In *Academy of Management Journal* 66 (4), pp. 1071–1101. DOI: 10.5465/amj.2020.1614.

Selbst, Andrew D.; Boyd, Danah; Friedler, Sorelle A.; Venkatasubramanian, Suresh; Vertesi, Janet (2019): Fairness and Abstraction in Sociotechnical Systems. In : Proceedings of the Conference on Fairness, Accountability, and Transparency. FAT* '19: Conference on Fairness, Accountability, and Transparency. Atlanta GA USA, 29 01 2019 31 01 2019. New York, NY, USA: ACM, pp. 59–68.

Shin, Donghee (2021): The effects of explainability and causability on perception, trust, and acceptance: Implications for explainable AI. In *International Journal of Human-Computer Studies* 146, p. 102551. DOI: 10.1016/j.ijhcs.2020.102551.

Shin, Donghee; Park, Yong Jin (2019): Role of fairness, accountability, and transparency in algorithmic affordance. In *Computers in Human Behavior* 98, pp. 277–284. DOI: 10.1016/j.chb.2019.04.019.

Shwartz-Ziv, Ravid; Armon, Amitai (2022): Tabular data: Deep learning is not all you need. In *Information Fusion* 81, pp. 84–90. DOI: 10.1016/j.inffus.2021.11.011.

Smith, Malcolm (2019): Research methods in accounting. 5. Auflage. Los Angeles u.a.: SAGE Puplications.

Somers, M. J.; Birnbaum, D.; Casal, J. (2021): Supervisor support, control over work methods and employee well-being: new insights into nonlinearity from artificial neural networks. In *International Journal of Human Resource Management* 32 (7), pp. 1620–1642. DOI: 10.1080/09585192.2018.1540442.

Speer, Andrew B. (2021): Empirical attrition modelling and discrimination: Balancing validity and group differences. In *Human Resource Management Journal*. DOI: 10.1111/1748-8583.12355.

Strohmeier, Stefan; Piazza, Franca (2015): Artificial Intelligence Techniques in Human Resource Management—A Conceptual Exploration. In Cengiz Kahraman, Sezi Çevik Onar (Eds.): Intelligent Techniques in Engineering Management, vol. 87. Cham: Springer International Publishing (Intelligent Systems Reference Library), pp. 149–172.

Tambe, Prasanna; Cappelli, Peter; Yakubovich, Valery (2019): Artificial Intelligence in Human Resources Management: Challenges and a Path Forward. In *California Management Review* 61 (4), pp. 15–42. DOI: 10.1177/0008125619867910.

Teodorescu, Mike; Morse, Lily; Awwad, Yazeed; Kane, Gerald (2021): Failures of Fairness in Automation Require a Deeper Understanding of Human-ML Augmentation. In *MIS Quarterly* 45 (3), pp. 1483–1500. DOI: 10.25300/MISQ/2021/16535.

Terwiesch, Christian (2023): Would chat GPT get a Wharton MBA. A prediction based on its performance in the operations management course. https://www.it-world.ru/upload/docs/lkfdgjeroip.pdf.

Time (2023): How ChatGPT Managed to Grow Faster Than TikTok or Instagram. https://time.com/6253615/chatgpt-fastest-growing/, updated on 11/24/2023, checked on 11/24/2023.

Turing, A. (1950): Computing Machinery and Intelligence. In *Mind* LIX (236), pp. 433–460. DOI: 10.1093/mind/LIX.236.433.

Tursunbayeva, Aizhan; Di Lauro, Stefano; Pagliari, Claudia (2018): People analytics—A scoping review of conceptual boundaries and value propositions. In *International Journal of Information Management* 43, pp. 224–247. DOI: 10.1016/j.ijinfomgt.2018.08.002.

Tursunbayeva, Aizhan; Pagliari, Claudia; Di Lauro, Stefano; Antonelli, Gilda (2021): The ethics of people analytics: risks, opportunities and recommendations. In *Personnel Review*. DOI: 10.1108/PR-12-2019-0680.

Vale, Daniel; El-Sharif, Ali; Ali, Muhammed (2022): Explainable artificial intelligence (XAI) post-hoc explainability methods: risks and limitations in non-discrimination law. In *AI and Ethics* 2 (4), pp. 815–826. DOI: 10.1007/s43681-022-00142-y.

Valizade, Danat; Schulz, Felix; Nicoara, Cezara (2024): Towards a Paradigm Shift: How Can Machine Learning Extend the Boundaries of Quantitative Management Scholarship? In *British Journal of Management* 35 (1), pp. 99–114. DOI: 10.1111/1467-8551.12678.

van den Broek, Elmira; Sergeeva, Anastasia; Huysman Vrije, Marleen (2021): When the Machine Meets the Expert: An Ethnography of Developing AI for Hiring. In *MIS Quarterly* 45 (3), pp. 1557–1580. DOI: 10.25300/MISQ/2021/16559.

Vargas, R.; Yurova, Y. V.; Ruppel, C. P.; Tworoger, L. C.; Greenwood, R. (2018): Individual adoption of HR analytics: a fine grained view of the early stages leading to adoption. In *International Journal of Human Resource Management* 29 (22), pp. 3046–3067. DOI: 10.1080/09585192.2018.1446181.

Vargas, Roslyn (2016): Adoption factors impacting human resource analytics among human resource professionals. US: ProQuest Information & Learning (Adoption factors impacting human resource analytics among human resource professionals., 76).

Venkatesh, Viswanath; Morris, Michael G.; Davis, Gordon B.; Davis, Fred D. (2003): User Acceptance of Information Technology: Toward a Unified View. In *MIS Quarterly* 27 (3), pp. 425–478. DOI: 10.2307/30036540.

Verdonk, Petra; Benschop, Yvonne W. M.; Haes, Hanneke C. J. M. de; Lagro-Janssen, Toine L. M. (2009): From gender bias to gender awareness in medical education. In *Advances in health sciences education : theory and practice* 14 (1), pp. 135–152. DOI: 10.1007/s10459-008-9100-z.

Vinuesa, Ricardo; Azizpour, Hossein; Leite, Iolanda; Balaam, Madeline; Dignum, Virginia; Domisch, Sami et al. (2020): The role of artificial intelligence in achieving the Sustainable Development Goals. In *Nature Communications* 11 (1), p. 233. DOI: 10.1038/s41467-019-14108-y.

Wang, Lijun; Zhou, Yu; Sanders, Karin; Marler, Janet H.; Zou, Yunqing (2024): Determinants of effective HR analytics Implementation: An In-Depth review and a dynamic framework for future research. In *Journal of Business Research* 170, p. 114312. DOI: 10.1016/j.jbusres.2023.114312.

Wang, Xin; Wang, Li; Zhang, Li; Xu, Xiaobo; Zhang, Weiyong; Xu, Yingcheng (2017): Developing an employee turnover risk evaluation model using case-based reasoning. In *Information Systems Frontiers* 19 (3), pp. 569–576. DOI: 10.1007/s10796-015-9615-9.

Yakusheva, Olga; Bang, James T.; Hughes, Ronda G.; Bobay, Kathleen L.; Costa, Linda; Weiss, Marianne E. (2022): Nonlinear association of nurse staffing and readmissions uncovered in machine learning analysis. In *Health services research* 57 (2), pp. 311–321. DOI: 10.1111/1475-6773.13695.

Yin, Robert K. (2013): Validity and generalization in future case study evaluations. In *Evaluation* 19 (3), pp. 321–332. DOI: 10.1177/1356389013497081.

Yuan, Shuai; Kroon, Brigitte; Kramer, Astrid (2021): Building prediction models with grouped data: A case study on the prediction of turnover intention. In *Human Resource Management Journal*. DOI: 10.1111/1748-8583.12396.

Zhou, Jianlong; Chen, Fang; Holzinger, Andreas (2022): Towards Explainability for AI Fairness. In Andreas Holzinger, Randy Goebel, Ruth Fong, Taesup Moon, Klaus-Robert Müller, Wojciech Samek (Eds.): xxAI - Beyond Explainable AI, vol. 13200. Cham: Springer International Publishing (Lecture Notes in Computer Science), pp. 375–386.

Zirar, Araz (2023): Can artificial intelligence's limitations drive innovative work behaviour? In *Review of Managerial Science* 17 (6), pp. 2005–2034. DOI: 10.1007/s11846-023-00621-4.