

Association for Information Systems

## AIS Electronic Library (AISeL)

---

PACIS 2023 Proceedings

Pacific Asia Conference on Information  
Systems (PACIS)

---

7-8-2023

# The Explanation Matters: Enhancing AI Adoption in Human Resource Management

Lorenz Baum

*Goethe University Frankfurt*, baum@wiwi.uni-frankfurt.de

Patrick Weber

*Goethe University Frankfurt*, weber@wiwi.uni-frankfurt.de

Laura-Marie Kolb

*Goethe University Frankfurt*, laura-marie.kolb@web.de

Follow this and additional works at: <https://aisel.aisnet.org/pacis2023>

---

### Recommended Citation

Baum, Lorenz; Weber, Patrick; and Kolb, Laura-Marie, "The Explanation Matters: Enhancing AI Adoption in Human Resource Management" (2023). *PACIS 2023 Proceedings*. 17.

<https://aisel.aisnet.org/pacis2023/17>

This material is brought to you by the Pacific Asia Conference on Information Systems (PACIS) at AIS Electronic Library (AISeL). It has been accepted for inclusion in PACIS 2023 Proceedings by an authorized administrator of AIS Electronic Library (AISeL). For more information, please contact [elibrary@aisnet.org](mailto:elibrary@aisnet.org).

# The Explanation Matters: Enhancing AI Adoption in Human Resource Management

*Completed Research Paper*

**Lorenz Baum**

Goethe University Frankfurt  
Theodor-W.-Adorno-Platz 4  
60323 Frankfurt am Main, Germany  
baum@wiwi.uni-frankfurt.de

**Patrick Weber**

Goethe University Frankfurt  
Theodor-W.-Adorno-Platz 4  
60323 Frankfurt am Main, Germany  
weber@wiwi.uni-frankfurt.de

**Laura-Marie Kolb**

Goethe University Frankfurt  
Theodor-W.-Adorno-Platz 4, 60323 Frankfurt am Main, Germany  
laura-marie.kolb@web.de

## Abstract

*Artificial intelligence (AI) has ubiquitous applications in companies, permeating multiple business divisions like human resource management (HRM). Yet, in these high-stakes domains where transparency and interpretability of results are of utmost importance, the black-box characteristic of AI is even more of a threat to AI adoption. Hence, explainable AI (XAI), which is regular AI equipped with or complemented by techniques to explain it, comes in. We present a systematic literature review of n=62 XAI in HRM papers. Further, we conducted an experiment among a German sample (n=108) of HRM personnel regarding a turnover prediction task with or without (X)AI-support. We find that AI-support leads to better task performance, self-assessment accuracy, and response characteristics toward the AI, and XAI, i.e., transparent models allow for more accurate self-assessment of one's performance. Future studies could enhance our research by employing local explanation techniques on real-world data with a larger and international sample.*

**Keywords:** Human resource management, AI adoption, Explainable artificial intelligence, Turnover prediction, Systematic Literature Review

## Introduction

As artificial intelligence (AI) continues to permeate every aspect of modern private and working lives, companies increasingly must manage AI (Berente et al., 2021). A recurring challenge of AI is its black-box characteristic, i.e., the fact that the input and output of an AI-system are observable, but the processing steps in between are not. Explainable AI (XAI) mitigates this challenge via transparent models or post-hoc explainability (Arrieta et al., 2020).

Information Systems academics increasingly engage in XAI research on application domains like HRM (Colace et al., 2019), finance (Weber, Carl, et al., 2023), and law (Bench-Capon et al., 2012). Within HRM, the importance of transparency, explainability, and interpretability is essential due to the highly consequential decisions involved; constituting prerequisites for the sustainable application and adoption of AI-systems (Janiesch et al., 2021; Mirbabaie et al., 2021). Therefore, these high-stakes decisions make HRM a domain with an intrinsic need for XAI on the one hand and, on the other hand, HRM decisions directly affect past, current, and future employees and, thus, are a vital contributor to a company's success (Noe et

al., 2020). Applications of AI in HRM span over multiple subareas, e.g., strategic planning, personnel search, and acquisition, personnel selection, administrative processing of HRM activities, communication with (potential) employees, development and implementation of training measures, employee evaluations, development of measures for employee retention, and evaluation of the potential of managers, and HRM personnel assign these subareas varying importance (Weber, 2023). The importance of explainability to create value in these subareas naturally varies, thus, subsequently, XAI is of varying importance for these subareas (Meske et al., 2022). For instance, in some subareas like development of measures for employee retention, decisions are highly consequential regarding their effect on the involved humans and financial aspects, while other subareas, like communication with employees, are less consequential. As Das and Rad (2020) correctly note, XAI outcomes currently cannot be blindly trusted. Additionally, challenges such as dealing with complex human dynamics, managing sensitive employee data, and addressing legal and ethical requirements are specific to HRM. This puts even more emphasis on the importance and adequacy of human-AI-collaboration in the field of HRM.

So far, research on XAI in HRM is sparse and researchers have been calling for applied XAI in fields like HRM (Langer & König, 2022). The study at hand aims to address this scarcity by employing an experiment conducted among a German national sample of  $n=108$  HRM personnel. We measure task performance, choice effort and difficulty, and attitude toward the information system (IS). Additionally, similar to Weber (2023) researching unrealistic optimism in AI in HRM, we compare performance and performance expectations (i.e., self-assessment) with and without AI-support. Furthermore, regarding the AI-support type, we divide the participants into three groups of which some receive additional XAI-support, and some do not. Thus, we want to answer the following three research questions (RQs) using their corresponding hypotheses (see subsection Hypotheses Development):

1. How does (X)AI-support affect task performance?
2. How does (X)AI-support affect self-assessment accuracy?
3. How does (X)AI-support affect further response characteristics toward an IS, such as choice effort, choice difficulty, and attitude toward the IS?

Interested readers from research may learn from our study that (X)AI-support increases task performance, improves self-assessment accuracy, and positively influences response characteristics toward the IS, i.e., the support system based on the (X)AI. Notably, for transparent XAI, we show more accurate self-assessment compared to black-box AI-support. This knowledge is particularly relevant for practitioners in HRM, as it highlights important considerations when adopting (X)AI systems in their domain. Further, our provided systematic literature review (SLR) results can serve as an entry point for future research.

This research paper is structured as follows: After this introductory section including the three RQs, the next section deals with the theoretical background by outlining XAI in HRM application including its adoption hurdles. Afterward, we will present the methodology of our study which includes developing the hypotheses in line with the research questions including a nomological network of the research and describing our study design. The section closes with demographic information about the participants next to the measures collected. In the results section, we present and visualize the main outcomes of our study before we conclude with theoretical and managerial implications, limitations of the study, and outlooks for future research in the last section.

## **Theoretical Background: Adoption Hurdles in AI in HRM Application**

AI refers to “machines that perform cognitive functions normally associated with the human mind, such as learning, interaction, and problem solving” (Raisch & Krakowski, 2020, p. 3). Research distinguishes between strong and weak AI (Russell et al., 2016; Turing, 1950). Strong AI, i.e., possessing (human-level) intelligence in a broad variety of fields, is fictitious at this point. However, instances of weak AI, i.e. (superhuman) intelligence in a narrowly defined field, are numerous, e.g., automated programming and interactive interpreters (Russell et al., 2016). These AI-systems have been trained to perform a specific task, but do not act intelligently beyond that. The skills required for this, such as abstract thinking and creativity, are only possessed by strong AI-systems (Russell et al., 2016; Turing, 1950).

AI models can be trained in several ways. For example, artificial neural networks (ANN) that mimic the human brain can be used (Russell et al., 2016). These consist of plexuses of neurons connected to varying degrees. The more complex an algorithm is, the more difficult its internal processes are for humans to

understand. Even simpler ANN already can consist of hundreds of weights of neurons, which embody the learned knowledge. Input and output of the model can be observed, but the strength of the connections of the inner layers is unknown, or due to a large number of connections and weights incomprehensible to humans. This makes it nearly impossible for humans to detect false neuron connections. This is the so-called black-box-characteristic of most AI, which allows humans to observe and interpret input and output data but does not allow an interpretation of the processing steps in between (Russell et al., 2016). However, as AI is also involved in more and more vital decisions, its results must be correct and comprehensible. AI, after all, does not have a human understanding of data. Algorithms cannot identify biased results. They only recognize the underlying patterns in a data set, but cannot evaluate them morally (Arrieta et al., 2020).

Methods from XAI offer a way to improve the interpretability of AI models by providing insights into processes and functions (Arrieta et al., 2020). The ultimate goal of these methods is to increase confidence in AI decisions by increasing the interpretability of AI models, which should enable and accelerate the adoption of AI in private and professional environments. XAI manifests in either transparent models or post-hoc explainability. Inherently interpretable models, that are interpretable without further due, are also called transparent (AI) models. A distinction is made between the degrees of algorithmic transparency, decomposability, and simulatability (Arrieta et al., 2020). As a representation of transparent XAI, a logistic regression assumes a linear connection between predictors and predicted variables, and uses the former to predict the dependent variables.

As mentioned above, most AI models suffer from the black-box characteristic. In these cases, humans cannot understand or decompose and simulate them, thus, the additional use of explanatory methods following the AI helps raise confidence. This form of explanation of an already existing or very complex AI is called post-hoc explainability, with multiple techniques including, e.g., simplification or visualization. The applicability of the different techniques in these categories differs depending on the machine learning algorithm used (Arrieta et al., 2020). Some methods are independent of the algorithm used and can be applied universally, i.e., model-agnostic explanation techniques. These techniques include Shapley Additive exPlanations (SHAP) and Local Interpretable Modelagnostic Explanations (LIME). SHAP is an example of the technique class of feature relevance explanation. It provides a score for the feature importance of every single prediction (Lundberg & Lee, 2017). LIME falls within the technique classes of explanation by simplification and local explanations. It explains a particular prediction by constructing a locally linear model around it (Ribeiro et al., 2016). Explanation techniques that can only be applied to certain algorithms are called model-specific, e.g., support vector machines, recurrent neural networks, or XGBoost. The latter combines multiple decision trees with weak prediction power over several iterations into a strong prediction model. Herein, the algorithm weighs the training data observations based on the previous iterations' error terms, i.e., the eponymous gradient boosting (Chen & Guestrin, 2016).

XAI is especially important in so-called high-risk environments with highly consequential decisions, such as HRM, finance, and law (Weber, Carl, et al., 2023). As HRM decisions directly affect past, current, or future human employees, and are a vital contributor to a company's success (Noe et al., 2020), AI-support of these decisions is in special need of XAI.

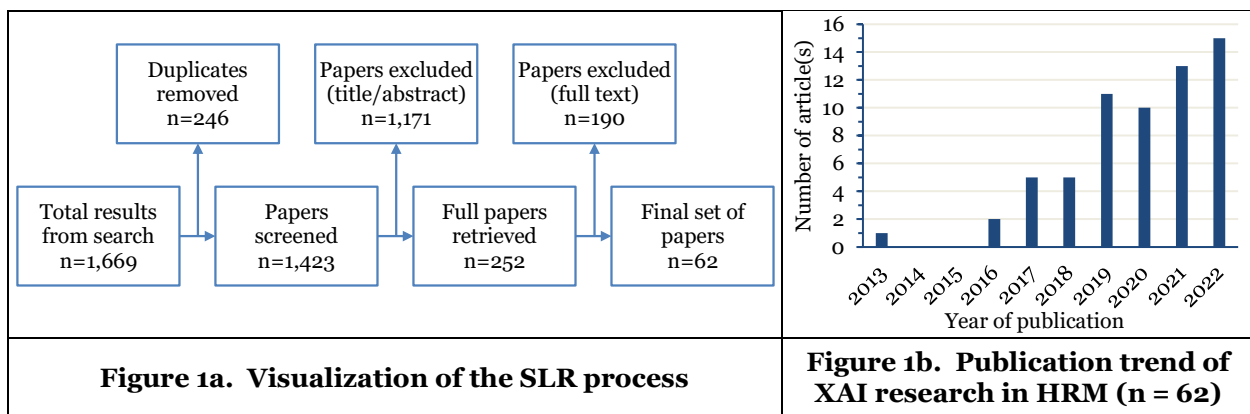
We employed an SLR to get a thorough overview of the state of research on XAI in HRM (Kitchenham et al., 2011), up to the year 2022. We incorporated multiple databases (JSTOR, Web of Science, AISEL, ScienceDirect, Google Scholar) and keywords<sup>1</sup> for XAI (Explainable Artificial Intelligence, erklärbare künstliche Intelligenz, explainable machine learning, explainability + artificial intelligence, Debugging, glass box, decision support system, XAI, explainability, transparency, actionable dashboard, explainable machine learning challenge, ChaLearn, machine learning, decision trees, deep residual networks, explainable AI) in combination with keywords for HRM (HRM, human capital, workforce management, Personalwesen, human resources, HR, human resource management, Personalmanagement, Recruiting, Recruitment, job interview, automatic job candidate screening, Video-CVs, CVs, apparent personality, candidate essay, personal selection, career recommendation, recommendation system, algorithmic job candidate screening, automatic recruitment, job mediator, job seeker, talent management, labor market, personality, algorithm-based resource management, LinkedVis, interview). This far-reaching search in the first step was necessary, to include as many potentially relevant articles as possible. Our search took place from June to December 2022. First, we removed 246 duplicates from our set. We reviewed relevant

---

<sup>1</sup> The keywords include English and German terms to maximize coverage of the SLR.

publications and screened their titles and abstracts (see Figure 1a). We employed the following rigid exclusion criteria to filter out irrelevant publications and ensure the results' relevance to the research goal: The paper deals (1) with HRM but not with XAI, (2) with XAI but not with HRM, (3) with AI and HRM but not with XAI, or (4) with XAI and HRM but does not focus XAI application in HRM. This step further reduced our set by 1,171 articles. Hence, we retrieved full texts for 252 papers and applied the aforementioned exclusion criteria again to further filter our result set. After this final step, our set consisted of 62 publications (see Appendix A). With this procedure, we ensured a thorough theoretical background to the manuscript at hand. We present excerpts of the SLR in the following paragraphs (see Figure 1a).

As Figure 1b shows, XAI research in HRM dates back as early as 2013 with an increase over the past years since 2016. The main barriers to AI adoption in HRM are the high complexity of HRM issues, data-related challenges, legal requirements, and the need for fairness and employee reactions (Tambe et al., 2019). Compared to human-led decision-making, applicants show lower trust, lower perceived fairness, and strong privacy concerns when selected by an algorithm (König & Langer, 2022). Another adoption hurdle might consist of a (perceived) trade-off between performance and trust. Current research discusses such thoughts and emphasizes being cautious regarding potential predictive performance losses (Sokol & Flach, 2020).



XAI can be usefully applied in all AI-systems of HRM to increase their transparency. However, this work focuses on the use of XAI in the prediction of employee turnover. The attrition of trained employees is a major challenge in organizations (Dachrodt et al., 2014). The success of an organization directly depends on its employees and their qualifications. Thus, employees are a crucial resource for the company (Noe et al., 2020). Consequently, a resigning employee not only causes costs arising from the rehiring process for that position but also manifests a knowledge loss for the company. Therefore, identifying the reasons for employee turnover and the means to prevent this is worthwhile for companies (Strohmeier & Piazza, 2015; Yuan et al., 2021). To this end, AI may examine data from former employees to predict the turnover probability of current employees. As these AI implementations often come with a black-box characteristic, XAI methods can be used to better understand automatically generated insights regarding the reasons why employees want to leave.

For example, Sekaran and Shanmugam (2022) apply approaches such as SHAP and LIME to determine the key factors influencing employee turnover using a gradient-boosting algorithm. The output of the influencing factors hardly differs between the two methods, which the authors interpret as a sign of the effectiveness of XAI methods. In our sample's first publication, Bostandjiev et al. (2013) introduce LinkedVis, a post-hoc explanation by visualization technique. LinkedVis implements natural language processing, i.e., automated, algorithmic retrieving of information from written texts (Russell et al., 2016), and entity resolution to present the user with professionals pursuing similar career paths in the professional social network LinkedIn. Based on these professionals, LinkedVis recommends job opportunities and companies. Herein, an interactive interface visualizes several aspects of the algorithm and allows users to manipulate profile item and social connection weights, thus serving as an explanatory mechanism of the underlying AI.

Langer et al. (2021) also state that a lack of explanations in human-computer interaction does not harm the perceived fairness of information, but does have a negative impact on the comprehensibility of the underlying decision process. According to Tsiakas and Murray-Rust (2022), the success of explanations

depends on several factors, such as the form of presentation, the frequency, the degree of transparency, and the form of the explanation. Another important aspect is the fit between the presentation of the explanation on the one hand and the context and the target audience on the other. Some authors even state that it is better to present users with *no* explanations at all than with the “wrong” ones (Langer et al., 2021).

## Method

### *Hypotheses Development*

Drawing on information processing theory (Atkinson & Shiffrin, 1968) as an overarching framework, we developed three hypotheses to investigate our research questions. Figure 2 visualizes the nomological network underlying these hypotheses and, thus, the study’s survey design. According to the RQs, we research the moderating implications of AI adoption (i.e., degree of AI- and XAI-support) on the three outcomes task performance, self-assessment, and response characteristics in the HRM setting.

#### **How does (X)AI-support affect task performance?**

According to information processing theory, task performance can be influenced by capabilities in terms of cognitive processes and long-term memory, and environmental input. Given a task (see next subsection for task description) a professional (i.e., HRM personnel) solves using an IS, the resulting task performance is mainly determined by the human skills (i.e., HRM skills) (Hunter, 1986). In addition, this IS might incorporate intelligent features, such as AI- or XAI-support which could affect the task performance as well, as the system can assist HRM personnel by augmenting their cognitive abilities and providing relevant insights. It is important to note that XAI-support implies AI-support, as XAI is a subcategory of AI (Arrieta et al., 2020). When comparing no intelligent support and AI-support, in general, we expect greater task performance when supported by an AI (Rai et al., 2019). Depending on the type of AI-support, i.e., type of environmental cue, we expect differences as well. Humans supported by an AI with black-box characteristics might achieve different performances than humans supported by a transparent or post-hoc explained AI (XAI-support). XAI allows HRM personnel to better understand the reasoning behind the AI’s recommendations and might lead to more informed decisions (Arrieta et al., 2020). Together, this leads to our first hypothesis H1: *The use of AI methods leads to a higher performance level than without AI (H1a), with higher performance being achieved when supported by XAI methods (H1b).*

Hence, we investigate the moderating effects of AI-support (H1a) and XAI-support (H1b) on individuals’ task performance. Personal HRM skills, which we cannot easily observe in our study, naturally affect task performance. By comparing a subject’s performance in a task without and with the support of an AI or XAI, the effect of HRM skills on performance is mitigated. This difference is our proxy for the effect of the support type on performance. AI-support should lead to an increase in objective task performance (H1a), whereby XAI-support should additionally increase the performance level when compared to mere AI-support (H1b).

#### **How does (X)AI-support affect self-assessment accuracy?**

Facing difficult tasks, people tend to overestimate their performance (Moore & Healy, 2008). Following Dunning (2011) and Atkinson and Shiffrin (1968), the accuracy of this self-assessment mainly depends on human expertise and experience (i.e., HRM skills). But additional information, i.e., environmental cues, about the task or the performance of others might also influence the personal self-assessment (Atkinson & Shiffrin, 1968). Hence, supporting humans through AI when there is no information about other humans’ performance could lead to more accurate self-assessments, as humans now have a basis for comparison. Following the same argument as above, the type of AI-support might as well entail changes to the accuracy of self-assessment (Kantack et al., 2022). Humans could benefit from additional details provided by transparent or post-hoc explained AI models, and thus experience an increase in accuracy or be less likely to overestimate their performance. These considerations lead to our second hypothesis H2: *The use of AI methods increases the self-assessment accuracy of one’s performance than without AI (H2a), with higher accuracy when supported by XAI methods (H2b).*

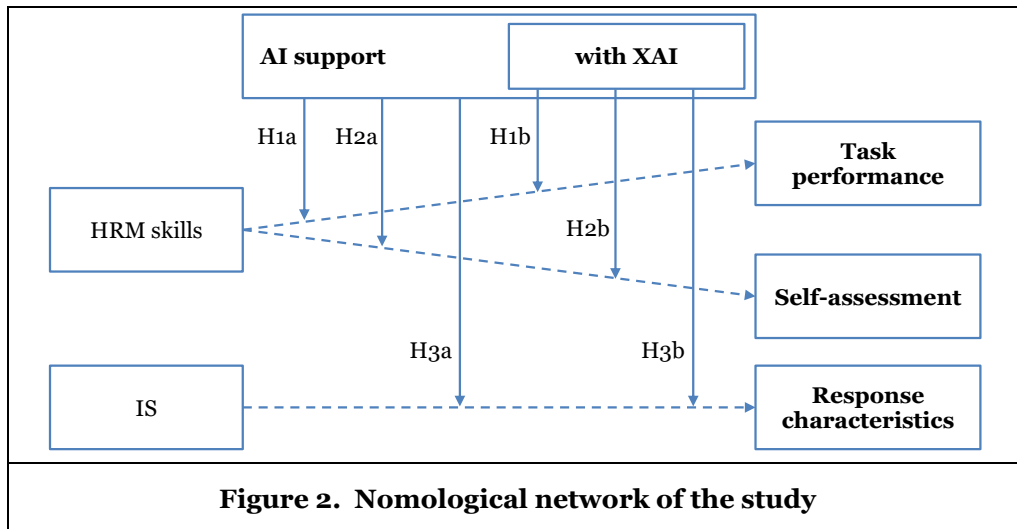
Again, we cannot reliably measure HRM skills. Contrarily, we rely on the comparison of the self-assessment between the scenarios with and without (X)AI-support to mitigate the main effect of the subject’s HRM skills. Comparing the (absolute) differences between task performance and self-assessment allows us to estimate the moderating effects of AI-support (H2a) regarding the accuracy (absolute difference) and level

of over-/underestimation of self-assessment (difference). Using the same measures, we can compare the effects of different support types (with and without XAI) on the quality of self-assessment. AI-support should lead to lower absolute differences (more accuracy) and less overestimation (H2a), and XAI-support should improve these measures further when compared to mere AI-support (H2b).

**How does (X)AI-support affect further response characteristics toward an IS, such as choice effort, choice difficulty, and attitude toward the IS?**

Finally, we want to investigate response characteristics toward the IS, as general idiosyncrasies of an IS naturally influence human response characteristics. Hence, we assume that not only the IS but also its features (i.e., (X)AI-support), both, as environmental cues, influence users’ opinions regarding the IS. While HRM is still the context, the response characteristics are closely related to the evaluation and perception of the IS itself as the relevant antecedent, as it represents the system and its features influencing users’ opinions. For this, we decided to employ three established measurement constructs to query the humans’ opinions about the IS (see Appendix B for details about the constructs): First, we adapted the evaluation costs scale by Heitmann et al. (2007) to our task to measure participants’ perceived choice effort when performing the task. Similar to Kelting et al. (2017), we employed the choice ease scale to assess the participants’ choice difficulty during the task. Lastly, the information value of the web site scale by Holzwarth et al. (2006) served as a measurement regarding the participants’ attitude toward the IS. We assume that through the support of AI, the perceived choice effort and difficulty should be reduced as the human possibly needs a lower cognitive capacity to work on the task. Differently, the human attitude toward the IS might increase when supported by an AI. Again, depending on the type of AI-support these effects might be further strengthened. Humans supported by a transparent or post-hoc explained AI (XAI-support) might report better response characteristics regarding the IS than humans supported by an AI with black-box characteristics due to the improved interpretability of the XAI-support. Together, this forms our third hypothesis H3: *The use of AI leads to better response characteristics toward the IS than without AI (H3a), with better response characteristics observed when supported by XAI methods (H3b).*

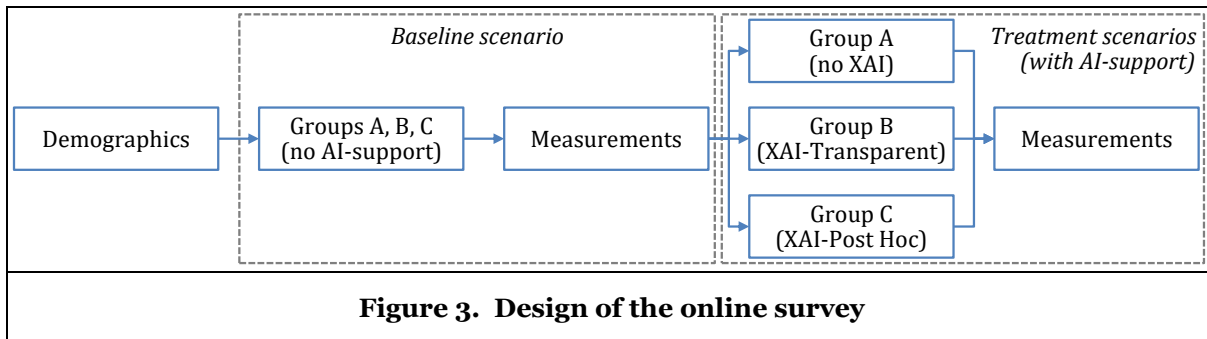
We refer to “better response characteristics” as reductions in choice effort and difficulty and increases in attitude toward the IS. By keeping the base IS characteristics the same and alternating only the type of support (no support or AI/XAI-support) we thus can mitigate the effects of base IS characteristics and measure the moderating effects of AI- or XAI-support by comparing those scenarios. AI-support is expected to ultimately lead to better response characteristics (H3a) and, by adding further interpretability, XAI-support should improve response characteristics even more when compared to mere AI-support (H3b). It is to be noted that though we assume HRM skills not affecting the response characteristics, potential effects by personal skills are as well mitigated by the use of the sequential design.



**Figure 2. Nomological network of the study**

### Study Design

To test the hypotheses, we conducted an online survey of HRM employees in Germany. To be able to compare our measures for different types of support, we chose a sequential design for the study. Figure 3 shows that after initial questions about sociodemographic data and control measurements (see Table 2), the participants solved an HRM-related task (turnover prediction) twice. In the baseline scenario, participants first performed the task without intelligent support, and afterward through the support of an AI or XAI. For this, the participants were randomly assigned to one of three groups A, B, and C. For the treatment scenarios, Group A solved the task with AI-support, Group B with the help of a transparent XAI, and Group C supported by a post-hoc XAI. Following each task, we measured task performance, self-assessment, and response characteristics. This design allows us to test H1a, H2a, and H3a using paired-samples tests on the whole sample. We then tested H1b, H2b, and H3b by comparing the groups using independent-samples tests.



**Figure 3. Design of the online survey**

For the study, we chose employee turnover prediction as the specific HRM use case and task, as it has sufficient complexity to be supported by AI. For the task, the participants were asked to sort five different employees according to their turnover probability (see Figure 4). We gathered the employee data from a publicly available, labeled dataset on Kaggle provided by IBM<sup>2</sup>, to present participants with a situation as realistic as possible. To this end, we trained an AI (logistic regression) to classify employees into risk groups according to their turnover probability based on their characteristics. To add further rigor to our study, we used these results as the output for all (X)AI predictions. Finally, we created four different dashboards, which we presented to the participants as our treatments. Figure 4 shows the basic dashboard for the baseline scenario which does not include any AI-support. As we surveyed HRM professionals in Germany, the dashboards are in German. Below the figures we present translations.

Name	Alter	Familienstand	Abteilung	Rolle	Grundgehalt	Überstunden	Reisen	Entfernung Wohnort	Betriebszugehörigkeit
A. Adam	32	Geschieden	Vertrieb	Vertriebsmitarbeiter	2,827.00 €	Nein	Nein	2	0
S. Schultze	32	Ledig	Forschung	Labortechniker	4,025.00 €	Nein	Nein	29	8
L. Lang	44	Verheiratet	Personalwesen	Personalwesen	5,743.00 €	Ja	Häufig	1	1
P. Pieper	18	Ledig	Forschung	Labortechniker	1,420.00 €	Nein	Selten	3	0
B. Boll	33	Geschieden	Vertrieb	Vertriebsleiter	8,380.00 €	Ja	Selten	1	9

**Figure 4. Dashboard without AI**

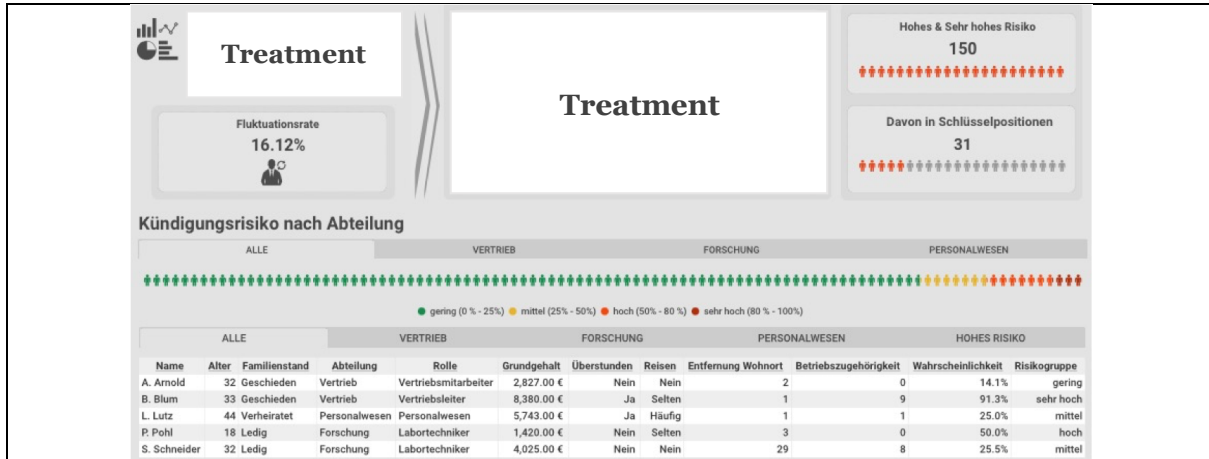
The table shows 5 different employees and their personnel data. This includes names, age, marital status, department, role, basic salary, overtime, travel activity, distance to their home, as well as years of service.

Figure 5 shows the general structure of the dashboards for the treatment scenarios. The task and the employee data presented here is the same as in the baseline scenario. Only the names of the employees were altered, and the turnover probability and risk groups were displayed. It is important to note that the remainder of the dashboard stayed the same across all groups; this especially includes the employee data which was not changed between the groups and the (X)AI predictions provided by the logistic regression to increase the reliability of the results and add further rigor and robustness. The white boxes in Figure 5 were the *only* detail replaced to treat the three groups.

<sup>2</sup> See <https://www.kaggle.com/code/thomaspmcg/ibm-employee-attrition> (last access May 30, 2023).



For Group A (AI-support with a black-box model), we described an ANN along with a general depiction of ANNs. We added a disclaimer mentioning the black-box characteristic of ANNs. Group B (XAI-support with a transparent model) was presented with support by a logistic regression including the calculated coefficients for the purpose of explanation. For Group C (XAI-support with post-hoc explanation), we explained the AI outcome using a chart of the 10 most important variables of a XGBoost model, which we trained on the IBM dataset. We ensured that these XGBoost explanations fit the presented results provided by the logistic regression (as mentioned earlier). Table 1 shows an overview of the three treatments.

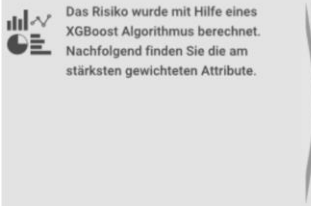



**Figure 5. Dashboard with AI**

The dashboard shows the basic framework for the three treatments with AI-support and is structured as follows:

- Top left: Fluctuation rate of the fictional company 16.12%
- Top right: Number of employees at high and very high risk of resigning (150), and the number of people in key positions thereof (31).
- Middle: Graphical representation indicating the distribution of turnover probabilities (low 0%-25%; medium 25%-50%; high 50%-80%; very high 80%-100%). The tabs can be used to switch between a representation of the whole company (“ALLE”) or single departments (sales, research, HRM)
- Bottom: Table with the same ten employee data columns as without AI (see Figure 4 above), only names were changed and supplemented by a calculated turnover probability and a risk group. The tabs above the table can be used to switch between an employee selection of the whole company (“ALLE”), single departments (sales, research, HRM) or those employees with a high risk.

Group	Treatment (translation below)	
Group A: AI-support with a black-box model		
	The risk was calculated with the help of an ANN (Multi-Layer Perceptron). The following provides a brief explanation.	An ANN is a “black-box” whose exact function cannot be reconstructed because of hidden layers.
Group B: XAI-support with a transparent model		
	The risk was calculated using logistic regression. The coefficients (weights) and the odds ratio (relative chance) of the individual attributes are listed in the following.	Structure of a logistic regression Formal definition: [formula] [variables, their coefficients, and odds ratios]

Group C: XAI-support with post-hoc explanation		
	The risk was calculated using an XGBoost algorithm. The following lists the most heavily weighted attributes.	Weights of the 10 most important variables [variables and their weights]
<b>Table 1. Overview of the AI and XAI treatments for Groups A, B, and C</b>		

### Participants

The study ran between August and November 2022. To recruit the participants, we used professional social networks and public information on job platforms approaching HRM personnel. In total, we noticed 249 responses to our survey, of which 114 participants completed the survey by reaching the last page. In this sample, before sanitization, we had a slight underrepresentation of Group B due to the randomization process. To add reliability and further rigor, we filtered participants using the employed attention check (5 exclusions) and the participation time (1 exclusion) to filter fast participants, as speeding indicates poor data quality (Greszki et al., 2015). We determined the lower boundary of acceptable participation time by subtracting one standard deviation (578 seconds) from the median participation time (805 seconds). In total, this process resulted in the elimination of 6 participants leaving the final sample at n=108.

Variable		Total	Standard Deviation
Age <sup>o</sup>	Mean (years)	35.5	10.1
Gender** <sup>o</sup>	female	64.8%	.48
	male	35.2%	
Level of education	Lower education/other	3.7%	
	High school diploma	19.4%	
	University/College degree	76.9%	
Field of education / background	Law	10.2%	
	Psychology	7.4%	
	Business	46.3%	
	other (e.g., politics, communication, pedagogy, sociology)	36.1%	
Uncertainty avoidance <sup>o</sup>	Mean (Likert scale 1-7)**	5.59	.77
Knowledge of AI <sup>o</sup>	Mean (Likert scale 1-7)**	2.85	1.43
Knowledge of XAI <sup>o</sup>	Mean (Likert scale 1-7)**	2.57	1.45
Need for cognitive closure <sup>o</sup>	Mean (Likert scale 1-6)**	3.93	.78

**Table 2. Descriptive statistics of the dataset (n=108)**

<sup>o</sup>No one chose the option “other” for gender.  
<sup>\*\*</sup>Scale from 1 “strongly disagree” to 7 (or 6) “strongly agree”.  
<sup>o</sup>No significant difference between treatment groups (randomization check).

The assignment to the groups was balanced with 36 people in each group A (AI-support with a black-box model), group B (XAI-support with a transparent model), and group C (XAI-support with post-hoc explanation). About 65% of the participants are female and the average age of the participants is about 36 years. In addition, participants have a high level of education as most participants (76.9%) have at least a bachelor’s degree, and about 59% have a master’s degree. Table 2 shows additional details about the field of education or background of the participants and four additional control measures. These four measures are especially interesting in the context of our AI-related task as they allow us to control for important differences between the treatment groups. We employed a construct to control for the participants uncertainty avoidance (Erdem et al., 2006), adapted the knowledge of the product class scale by Alavi et al. (2016) to measure AI and XAI knowledge, and used a measurement construct to assess the participants’ need for cognitive closure (Kruglanski et al., 1993).

To assess the success of the randomization process, independent-samples Kruskal-Wallis tests show that the treatment groups are equal with respect to participation time ( $\chi^2(2)=2.43$ ,  $p=.26$ ), gender ( $\chi^2(2)=4.18$ ,  $p=.12$ ), age ( $\chi^2(2)=3.05$ ,  $p=.22$ ), uncertainty avoidance ( $\chi^2(2)=1.03$ ,  $p=.60$ ), knowledge of AI ( $\chi^2(2)=.53$ ,  $p=.77$ ) and XAI ( $\chi^2(2)=.25$ ,  $p=.88$ ), and the need for cognitive closure ( $\chi^2(2)=4.01$ ,  $p=.13$ ). Thus, the randomization was successful. Ultimately, we checked for the presence of common method bias (CMB) in our sample using Harman's one-factor test. For both, principal axis factoring and principal component factoring, the findings revealed that more than one factor was present in the data with the explained variances being below the critical value of 50% (12.9% and 16.3%). This indicates that there is little chance of CMB interfering with our study's findings (Podsakoff et al., 2003).

## Measures

Participants performed the following task: sorting employees visualized in the dashboard according to their turnover probability. The person with the highest probability was to be assigned to 1st place. Then, we used the correct order based on the IBM dataset to count the number of pairs of two employees which the participants correctly placed relative to each other. With five employees to sort, this yields a maximum of  $\binom{5}{2}=10$  and a minimum of 0 correct pairs. We told the participants how this task performance is calculated using an illustrative example. After solving the tasks, we asked participants to assess their estimated performance using the same measure, enabling us to compare participants' objective and subjective performance. Table 3 summarizes the results regarding task performance, self-assessment, and response characteristics. Additionally, we calculated the difference between task performance and self-assessment for each scenario and participant to assess the accuracy regarding self-assessment, e.g., over-/underestimation, and the absolute difference between task performance and self-assessment accounting for the accuracy of the self-assessment. In the following section, we will use these results to answer the hypotheses for this study.

Measure (Mean (Standard Deviation))	No AI-support (n=108)	Total (n=108)	With AI-support		
			Group A (no XAI) (n=36)	Group B (XAI-Transparent) (n=36)	Group C (XAI-Post Hoc) (n=36)
Task performance (# of correctly assigned pairs)	4.3 (1.8)	7.2 (2.6)	7.4 (2.5)	6.9 (2.8)	7.3 (2.6)
Self-assessment (estimated # of correctly assigned pairs)	5.3 (1.8)	5.9 (2.1)	5.5 (2.1)	6.4 (2.1)	5.8 (2.1)
Difference (Task performance - Self- assessment; difference of # of pairs)	-1.1 (2.7)	1.3 (2.7)	1.9 (2.7)	.5 (2.3)	1.5 (3.0)
Absolute difference (Task performance - Self-assessment; absolute difference of # of pairs)	2.3 (1.7)	2.4 (1.8)	2.6 (1.9)	1.8 (1.6)	2.9 (1.7)
Choice effort*	4.5 (1.0)	3.9 (1.0)	4.0 (1.1)	4.0 (1.0)	3.7 (1.0)
Choice difficulty**	3.6 (1.3)	3.1 (1.4)	3.2 (1.4)	2.8 (1.5)	3.2 (1.4)
Attitude toward the IS*	4.1 (1.2)	4.5 (1.4)	5.0 (1.2)	4.1 (1.5)	4.5 (1.3)

**Table 3. Measurements for the treatment groups.**

\*Scale from 1 “completely disagree” to 7 “completely agree”.

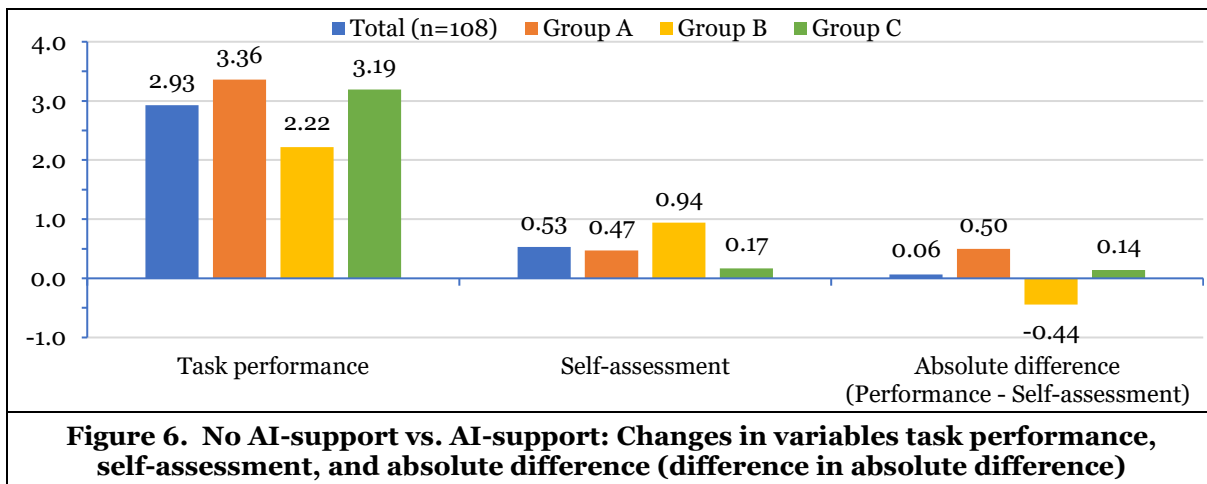
\*\*Scale from 1 “not at all” to 7 “extreme”.

## Results

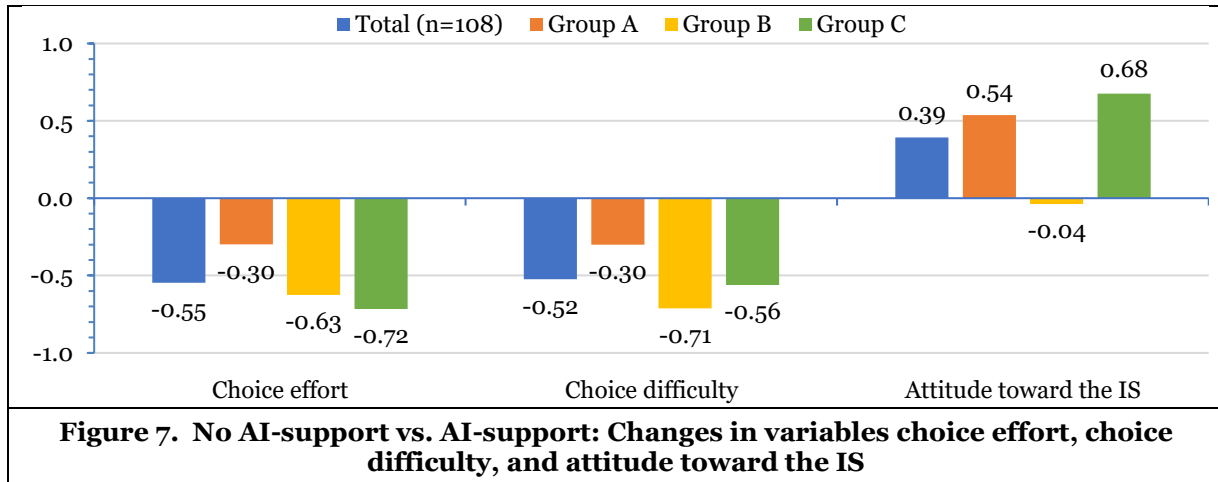
With  $n=108$  participants in total and  $n=36$  participants per treatment group, the use of parametric tests is possible ( $n>30$ ). Hence, we used paired-samples t-tests to investigate H1a. The use of AI improved the number of correctly matched pairs by 2.9 pairs to an average of 7.2. The t-test on an aggregated level shows that the difference in task performance between the scenario without AI and the scenarios with AI is highly significant ( $t=10.2$ ,  $p<.001$ ). Additionally, we tested whether this effect persists individually within the three treatment groups. These t-tests also reveal significant performance improvements, regardless of the type of treatment ( $p<.001$  for all three groups,  $n=36$ ). Thus, we find strong support for H1a, as task performance improves significantly with the support of AI. Figure 6 shows these improvements for all groups. When

comparing absolute task performance (Table 3) and performance improvement, there seems to be no higher performance for XAI-supported groups. Contrarily, there might be a slight advantage for the black-box-supported group, compared to the Group B. Analysis of variance (ANOVA) shows that there is no significant difference between the groups regarding absolute task performance ( $F(2,105)=.32, p=.73$ ) and performance improvement ( $F(2,105)=1.55, p=.22$ ). We, therefore, must reject H1b.

Looking at Figure 6, self-assessment seems to have increased throughout the sample (by 0.5 pairs). But, more interestingly, the difference between performance and self-assessment changed its sign. For the baseline scenario participants systematically overestimated their performance and with AI-support they underestimated it. A paired-samples t-test confirms this highly significant change in the level of self-assessment ( $t=8.3, p<.001$ ). This result also holds true for all individual groups ( $p<.01$  for each group,  $n=36$ ). Therefore, we can confirm H2a regarding the accuracy of self-assessment. Contrarily, absolute differences between performance and self-assessment did not improve (getting smaller) when supported by an AI ( $t=.27, p=.79$ ). Looking at the right side of Table 3, it appears that both, difference and absolute difference are lower for the group supported by the transparent AI compared to the other groups. In fact, the ANOVA confirms weakly significant differences in differences ( $F(2,105)=2.48, p=.09$ ) and significant differences in absolute differences ( $F(2,105)=3.90, p=.02$ ) between the groups. Post-hoc-tests with Bonferroni correction confirm significant differences in absolute differences between Groups B and C, with Group B achieving higher accuracy ( $p=.03$ ). Thus, it appears that the support by transparent AI models converges the self-assessment to the actual performance, resulting in a more accurate self-assessment. This positive effect is damped by the slightly lower change in task performance for this group (not significant, see above). In summary, we only partly find evidence for H2b, i.e., for the group supported by the transparent XAI model regarding the accuracy of self-assessment.



For our third hypothesis, Figure 7 suggests that all three response characteristics improve when supported by an AI. The paired-samples t-tests confirm this observation. There are highly significant decreases in perceived choice effort ( $t=-4.45, p<.001$ ) and in choice difficulty ( $t=-3.29, p<.01$ ) and a highly significant increase in attitude toward the IS ( $t=2.87, p<.01$ ). Thus, we confirm H3a for the mere difference between no AI-support and AI-support. Individually within the groups, not all effects hold (choice effort is significant for Groups B and C, the choice difficulty is slightly significant for Groups B and C, and attitude toward the IS is significant for Groups A and C). In summary, for choice effort and difficulty the improvement can be also confirmed within the XAI-supported groups. A possible reason for the missing improvement in attitude for Group B could be the participants' lack of knowledge about how linear regressions work. Thus, the explanations provided might not be perceived as useful. A Spearman correlation test revealed that participants who achieved higher performance also showed better response characteristics ( $p<.01$ ). This also holds for Group B's attitude toward the IS ( $r_s=.69, p<.001$ ). Looking at absolute levels or changes in choice effort and difficulty for the treatment groups, there seems to be no difference between the groups. This finding is confirmed by the ANOVA (for all  $p>.1$ ). Only attitude toward the IS shows significant differences between the groups for absolute levels ( $F(2,105)=3.49, p=.03$ ) and changes ( $F(2,105)=2.63, p=.08$ ). A post-hoc-test with Bonferroni correction confirms the significantly lower attitude toward the IS for Group B when compared to Group A. In summary, we cannot confirm H3b.



## Conclusion

With the present study, we contribute to research in multiple ways. We summarize our found hypotheses support in Table 4. First, we apply XAI in the field of HRM answering researchers’ calls (e.g., Langer & König, 2022). Through this, we provide an investigation of (X)AI-support on task performance, self-assessment, and further response characteristics. Particularly, we find that through AI-support participants performed 67% better compared to no support (about 3 correct pairs more, see H1a). Additionally, through AI-support the systematic overestimation in difficult tasks, which we observed as well (Moore & Healy, 2008), was corrected toward an underestimation of one’s task performance (see H2a). The lack of performance increment with XAI-support compared to black-box AI-support might be outweighed for transparent models as, although their support might lead to a slightly lower increase in performance, with them the self-assessment is more realistic and accurate. This is especially beneficial for high-stakes domains like HRM where final decisions usually are made by humans which thus can better rely on explanations. Ultimately, also response characteristics improved with AI-support by about 0.5 points for choice effort and difficulty, and about 0.4 points for attitude toward the IS (see H3a). Here, XAI-support shows slightly greater improvements for choice effort and difficulty. Second, the sequential design and the results of our study could guide future research investigating the topic of (X)AI in different domains, based on different tasks or task complexities, or in different populations. Our results show that the way of combining objective performance with self-assessment can lead to interesting results regarding the adoption of (X)AI in a domain. Third, while research has already addressed areas of AI in Information Systems and Management (e.g., Abdel-Karim et al., 2021; Martin, 2019; Rai et al., 2019; Raisch & Krakowski, 2020), HRM (e.g., König & Langer, 2022; Tambe et al., 2019; Weber, 2023), or of XAI in general (e.g., Arrieta et al., 2020; Rosenfeld & Richardson, 2019; Sokol & Flach, 2020), so far, there is no systematic overview available regarding XAI in HRM, in particular. With our study and the conducted SLR, we provide an easy-to-follow, low-threshold, comprehensive overview that aggregates the still scattered research.

The present study also adds to the understanding of XAI as a form of information asymmetry (IA) reduction. When employing AI-systems, IA, i.e., a state of unequally distributed information in a mutual (planned) negotiation (Akerlof, 1970) may arise between the AI and the person using it, especially due to the above-mentioned black-box characteristic. In this situation, XAI may inform the user, thus, (partly) alleviating the introduced IAs. In the reported study, we assume IAs to be higher in the AI than in the XAI scenarios, as the latter explains the model’s inner mechanisms to the user.

	Task performance	Self-assessment	Response characteristics
<b>With AI-support</b>	✓ (H1a)	✓ (H2a)	✓ (H3a)
<b>With XAI-Support</b>	× (H1b)	(✓) (H2b)	× (H3b)

**Table 4. Overview of Hypotheses with their support**

Regarding managerial implications, first, the paper highlights important points that companies should consider when implementing AI-systems in HRM. The bare use of XAI methods does not necessarily lead to an increase in the attitude toward the IS (see Figure 7 - Group B), and, thus, might hinder adoption processes. Each of the presented (X)AI methods has multiple implications regarding task performance, choice effort, and difficulty, that companies need to balance. Therefore, the application context should always be considered, as well as the needs of the users. Second, for interested readers from HRM practice, this study may serve as an overview to inform themselves regarding the implementation possibilities of XAI in their business division. Our SLR provides an entry point for implementing XAI in HRM.

Although this work provides new insights into the effectiveness of XAI methods in the field of HRM, it is subject to some limitations that must be considered when interpreting the results. The underlying dataset is a simulated dataset and thus does not allow fully realistic training of AI. Thus, the displayed turnover probabilities may also differ from a real data set. Since the dashboards for this work should be comparable, only global explanatory methods were used. However, HRM managers are not only concerned with the global view of their workforce but also with individual employees. Therefore, our focus on global explanatory methods limits our understanding of the potential benefits of local explanation techniques, which may be important in HRM decision-making processes. The group of participants is relatively small with 108 participants and is limited to HRM employees only. Increasing the sample size would enhance the statistical power and generalizability of our results. Contrarily, replicating the study in a more controlled environment, such as a laboratory, could give deeper insights compared to an online experiment like ours. Additionally, controlling for participants' objective AI literacy (Weber, Pinski, et al., 2023) as it may influence the perception of XAI, or their level of experience within the HRM context, e.g., via occupation tenure, might lead to further insights. The survey's specific focus on employee turnover prediction may limit the generalizability of our results to other areas within HRM. Besides this, though mitigated through the investigation of differences, treatment order effects might affect our results, as participants always first performed the task without AI, and then with (X)AI-support. However, further investigations could consider alternative designs, e.g., adding a control group repeating the first task to reflect upon order effects, or counterbalancing to minimize such effects.

The work points to further issues that should be addressed in future research. In doing so, further experiments would be useful to better understand the cognitive load due to the explanations. It is important to identify which explanation form represents the optimal point between additional benefit, mental effort, and time expenditure for HRM personnel. Future researchers may replicate our study with a different task, a larger and international set of participants, or observe (X)AI effects in HRM over time. As mentioned above, in further studies there should also be a focus on local explanation methods, as these could be a useful addition, especially in HRM use cases. Also, investigating different quality levels of the explanations or the role of cognitive load in the context of XAI adoption are fruitful avenues for future research, next to researching different kinds of XAI-support, e.g., transparent models like decision trees or post-hoc explanations via individual conditional expectation plots. Furthermore, extending the investigation of XAI methods to other business domains beyond HRM would provide a broader understanding of their potential benefits.

## References

- Abdel-Karim, B. M., Pfeuffer, N., & Hinz, O. (2021). Machine Learning in Information Systems—a Bibliographic Review and Open Research Issues. *Electronic Markets*, 31(3), 643–670. <https://doi.org/10.1007/s12525-021-00459-2>
- Akerlof, G. A. (1970). Quality uncertainty and the market mechanism. *Quarterly Journal of Economics*, 84(3), 488–500.
- Alavi, S., Wieseke, J., & Guba, J. H. (2016). Saving on discounts through accurate sensing—salespeople's estimations of customer price importance and their effects on negotiation success. *Journal of Retailing*, 92(1), 40–55.
- Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., Garcia, S., Gil-Lopez, S., Molina, D., Benjamins, R., Chatila, R., & Herrera, F. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, 58, 82–115. <https://doi.org/10.1016/j.inffus.2019.12.012>

- Atkinson, R. C., & Shiffrin, R. M. (1968). Human Memory: A Proposed System and its Control Processes. In *Psychology of Learning and Motivation* (Vol. 2, pp. 89–195). Elsevier. [https://doi.org/10.1016/S0079-7421\(08\)60422-3](https://doi.org/10.1016/S0079-7421(08)60422-3)
- Bench-Capon, T., Araszkievicz, M., Ashley, K., Atkinson, K., Bex, F., Borges, F., Bourcier, D., Bourguine, P., Conrad, J. G., Francesconi, E., Gordon, T. F., Governatori, G., Leidner, J. L., Lewis, D. D., Loui, R. P., McCarty, L. T., Prakken, H., Schilder, F., Schweighofer, E., ... Wyner, A. Z. (2012). A history of AI and Law in 50 papers: 25 years of the international conference on AI and Law. *Artificial Intelligence and Law*, 20(3), 215–319. <https://doi.org/10.1007/s10506-012-9131-x>
- Berente, N., Gu, B., Recker, J., & Santhanam, R. (2021). Managing artificial intelligence. *MIS Quarterly*, 45(3), 1433–1450.
- Bostandjiev, S., O'Donovan, J., & Höllerer, T. (2013). LinkedVis: Exploring social and semantic career recommendations. *Proceedings of the 2013 International Conference on Intelligent User Interfaces*, 107–116.
- Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794. <https://doi.org/10.1145/2939672.2939785>
- Colace, F., De Santo, M., Lombardi, M., Mercurio, F., Mezzanzanica, M., & Pascale, F. (2019). *Towards labour market intelligence through topic modelling*.
- Dachrodt, H.-G., Koberski, W., Engelbert, V., & Dachrodt, G. (2014). *Praxishandbuch Human Resources*. Springer Fachmedien Wiesbaden.
- Das, A., & Rad, P. (2020). Opportunities and challenges in explainable artificial intelligence (xai): A survey. *ArXiv Preprint ArXiv:2006.11371*.
- Dunning, D. (2011). The Dunning–Kruger effect: On being ignorant of one's own ignorance. In *Advances in Experimental Social Psychology* (Vol. 44, pp. 247–296). Elsevier. <https://doi.org/10.1016/B978-0-12-385522-0.00005-6>
- Erdem, T., Swait, J., & Valenzuela, A. (2006). Brands as Signals: A Cross-Country Validation Study. *Journal of Marketing*, 70(1), Article 1. <https://doi.org/10.1509/jmkg.70.1.034.qxd>
- Greszki, R., Meyer, M., & Schoen, H. (2015). Exploring the effects of removing “too fast” responses and respondents from web surveys. *Public Opinion Quarterly*, 79(2), 471–503. <https://doi.org/10.1093/poq/nfu058>
- Heitmann, M., Lehmann, D. R., & Herrmann, A. (2007). Choice Goal Attainment and Decision and Consumption Satisfaction. *Journal of Marketing Research*, 44(2), Article 2. <https://doi.org/10.1509/jmkr.44.2.234>
- Holzwarth, M., Janiszewski, C., & Neumann, M. M. (2006). The Influence of Avatars on Online Consumer Shopping Behavior. *Journal of Marketing*, 70(4), Article 4. <https://doi.org/10.1509/jmkg.70.4.019>
- Hunter, J. E. (1986). Cognitive ability, cognitive aptitudes, job knowledge, and job performance. *Journal of Vocational Behavior*, 29(3), 340–362. [https://doi.org/10.1016/0001-8791\(86\)90013-8](https://doi.org/10.1016/0001-8791(86)90013-8)
- Janiesch, C., Zschech, P., & Heinrich, K. (2021). Machine Learning and Deep Learning. *Electronic Markets*, 31(3), 685–695. <https://doi.org/10.1007/s12525-021-00475-2>
- Kantack, N., Cohen, N., Bos, N., Lowman, C., Everett, J., & Endres, T. (2022). Instructive artificial intelligence (AI) for human training, assistance, and explainability. In T. Pham, L. Solomon, & M. E. Hohil (Eds.), *Artificial Intelligence and Machine Learning for Multi-Domain Operations Applications IV* (p. 5). SPIE. <https://doi.org/10.1117/12.2618616>
- Kelting, K., Duhachek, A., & Whitler, K. (2017). Can copycat private labels improve the consumer's shopping experience? A fluency explanation. *Journal of the Academy of Marketing Science*, 45(4), Article 4. <https://doi.org/10.1007/s11747-017-0520-2>
- Kitchenham, B., Budgen, D., & Brereton, O. P. (2011). Using Mapping Studies as the Basis for Further Research—a Participant-Observer Case Study. *Information and Software Technology*, 53(6), 638–651. <https://doi.org/10.1016/j.infsof.2010.12.011>
- König, C. J., & Langer, M. (2022). Machine learning in personnel selection. In *Handbook of Research on Artificial Intelligence in Human Resource Management* (pp. 149–167). Edward Elgar Publishing.
- Kruglanski, A. W., Webster, D. M., & Klem, A. (1993). *Motivated Resistance and Openness to Persuasion in the Presence or Absence of Prior Information*. 17.
- Langer, M., Baum, K., König, C. J., Hähne, V., Oster, D., & Speith, T. (2021). Spare me the details: How the type of information about automated interviews influences applicant reactions. *International Journal of Selection and Assessment*, 29(2), Article 2.



- Langer, M., & König, C. (2022). Explainability of artificial intelligence in human resources. *Handbook of Research on Artificial Intelligence in Human Resource Management*, 285–302.
- Lundberg, S. M., & Lee, S.-I. (2017). A Unified Approach to Interpreting Model Predictions. *Advances in Neural Information Processing Systems*, 30. <https://doi.org/10.5555/3295222.3295230>
- Martin, K. (2019). Designing Ethical Algorithms. *MIS Quarterly Executive*, 18(2), 129–142. <https://doi.org/10.2139/ssrn.3056692>
- Meske, C., Bunde, E., Schneider, J., & Gersch, M. (2022). Explainable Artificial Intelligence: Objectives, Stakeholders, and Future Research Opportunities. *Information Systems Management*, 39(1), 53–63. <https://doi.org/10.1080/10580530.2020.1849465>
- Mirbabaie, M., Brünker, F., Möllmann Frick, N. R. J., & Stieglitz, S. (2021). The Rise of Artificial Intelligence – Understanding the AI Identity Threat at the Workplace. *Electronic Markets*, 32, 73–99. <https://doi.org/10.1007/s12525-021-00496-x>
- Moore, D. A., & Healy, P. J. (2008). The trouble with overconfidence. *Psychological Review*, 115(2), 502–517. <https://doi.org/10.1037/0033-295X.115.2.502>
- Noe, R. A., Hollenbeck, J., Gerhart, B., & Wright, P. (2020). *Fundamentals of Human Resource Management* (8th ed.). Boston, Mass. McGraw-Hill.
- Podsakoff, P. M., MacKenzie, S. B., Lee, J.-Y., & Podsakoff, N. P. (2003). Common method biases in behavioral research: A critical review of the literature and recommended remedies. *Journal of Applied Psychology*, 88(5), 879.
- Rai, A., Constantinides, P., & Sarker, S. (2019). Next Generation Digital Platforms: Toward Human-AI Hybrids. *MIS Quarterly*, 43(1), iii–ix.
- Raisch, S., & Krakowski, S. (2020). Artificial Intelligence and Management: The Automation-Augmentation Paradox. In *Academy of Management Review*. <https://doi.org/10.5465/2018.0072>
- Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). “Why Should I Trust You?”: Explaining the Predictions of Any Classifier. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1135–1144. <https://doi.org/10.1145/2939672.2939778>
- Rosenfeld, A., & Richardson, A. (2019). Explainability in Human-Agent Systems. *Autonomous Agents and Multi-Agent Systems*, 33(6), 673–705. <https://doi.org/10.1007/s10458-019-09408-y>
- Russell, S. J., Norvig, P., Davis, E., & Edwards, D. (2016). *Artificial intelligence: A modern approach* (Third edition, Global edition). Pearson.
- Sekaran, K., & Shanmugam, S. (2022). Interpreting the Factors of Employee Attrition using Explainable AI. *2022 International Conference on Decision Aid Sciences and Applications (DASA)*, 932–936.
- Sokol, K., & Flach, P. (2020). Explainability fact sheets: A framework for systematic assessment of explainable approaches. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 56–67. <https://doi.org/10.1145/3351095.3372870>
- Strohmeier, S., & Piazza, F. (2015). Artificial intelligence techniques in human resource management—A conceptual exploration. In *Intelligent techniques in engineering management* (Vol. 87, pp. 149–172). Springer. [https://doi.org/10.1007/978-3-319-17906-3\\_7](https://doi.org/10.1007/978-3-319-17906-3_7)
- Tambe, P., Cappelli, P., & Yakubovich, V. (2019). Artificial Intelligence in Human Resources Management: Challenges and a Path Forward. *California Management Review*, 61(4), Article 4. <https://doi.org/10.1177/0008125619867910>
- Tsiakas, K., & Murray-Rust, D. (2022). Using human-in-the-loop and explainable AI to envisage new future work practices. *The 15th International Conference on Pervasive Technologies Related to Assistive Environments*, 588–594. <https://doi.org/10.1145/3529190.3534779>
- Turing, A. M. (1950). Computing Machinery and Intelligence. *Mind*, 59(236), 433–460.
- Weber, P. (2023). Unrealistic Optimism Regarding Artificial Intelligence Opportunities in Human Resource Management. *International Journal of Knowledge Management (IJKM)*, 19(1), 1–19. <https://doi.org/10.4018/IJKM.317217>
- Weber, P., Carl, K. V., & Hinz, O. (2023). Applications of Explainable Artificial Intelligence in Finance—A systematic review of Finance, Information Systems, and Computer Science literature. *Management Review Quarterly*. <https://doi.org/10.1007/s11301-023-00320-0>
- Weber, P., Pinski, M., & Baum, L. (2023). Towards an Objective Measurement of AI Literacy. *PACIS 2023 Proceedings*. Pacific Asia Conference on Information Systems 2023, Nanchang, China.
- Yuan, S., Kroon, B., & Kramer, A. (2021). Building prediction models with grouped data: A case study on the prediction of turnover intention. *Human Resource Management Journal*.



## Acknowledgments

This work has been funded by the German Research Foundation (DFG) within the Collaborative Research Center (CRC) 1053 MAKI.

## Appendix A: Complete XAI in HRM Sample with n=62 Papers

Author(s)	Title	Year
Alkan et al.	Where can my career take me? harnessing dialogue for interactive career goal recommendations	2019
Alkan et al.	Opportunity team builder for sales teams	2018
Arakawa & Yakura	Human-AI communication for human-human communication: Applying interpretable unsupervised anomaly detection to executive coaching	2022
Bañeres Besora & Conesa Caralt	A life-long learning recommender system to promote employability	2017
Bankins	The ethical use of artificial intelligence in human resource management: a decision-making framework	2021
Barrak et al.	Toward a traceable, explainable, and fairJD/Resume recommendation system	2022
Berger & Müller	Back to basics: Explainable AI for adaptive serious games	2021
Bostandjiev et al.	LinkedVis: exploring social and semantic career recommendations	2013
Campion et al.	Initial investigation into computer scoring of candidate essays for personnel selection.	2016
Chan	AI employment decision-making: integrating the equal opportunity merit principle and explainable AI	2022
Charleer et al.	Supporting job mediator and job seeker through an actionable dashboard	2019
Chhatwal et al.	Explainable text classification in legal document review a case study of explainable predictive coding	2018
Cho et al.	Toward Effective IT Services in Defence Talent Management Platform	2020
Choi et al.	A Study of the Classification of IT Jobs Using LSTM and LIME	2020
Chowdhury et al.	Embedding transparency in artificial intelligence machine learning models: managerial implications on predicting and explaining employee turnover	2022
Colace et al.	Towards labour market intelligence through topic modelling	2019
Delecraz et al.	Transparency and Explainability of a Machine Learning Model in the Context of Human Resource Management	2022
Doornenbal et al.	Opening the black box: Uncovering the leader trait paradigm through machine learning	2022
Escalante et al.	Design of an explainable machine learning challenge for video interviews	2017
Escalante et al.	Modeling, recognizing, and explaining apparent personality from videos	2020
Escalante et al.	Explaining first impressions: Modeling, recognizing, and explaining apparent personality from videos	2018
Fleiß et al.	Explainability and the intention to use AI-based conversational agents.	2020
Gonzalez et al.	“Where’s the IO?” Artificial intelligence and machine learning in talent management systems	2019
Goretzko & Israel	Pitfalls of Machine Learning-Based Personnel Selection	2021
Gucluturk et al.	Visualizing apparent personality analysis with deep residual networks	2017
Guleria & Sood	Explainable AI and machine learning: performance evaluation and explainability of classifiers on educational data mining inspired career counseling	2022
Gutiérrez et al.	Explaining and exploring job recommendations: a user-driven approach for interacting with knowledge-based job recommender systems	2019
He et al.	Interactive recommender systems: A survey of the state of the art and future research challenges and opportunities	2016
Heimerl et al.	“GAN I hire you?” –A System for Personalized Virtual Job Interview Training	2022
Jain et al.	Explaining and predicting employees’ attrition: a machine learning approach	2020
Jenkins et al.	Predicting success in United States Air Force pilot training using machine learning techniques	2022
Juvitayapun	Employee Turnover Prediction: The impact of employee event features on interpretable machine learning methods	2021
Kaya et al.	Multi-modal score fusion and decision trees for explainable automatic job candidate screening from video cvs	2017
Kaya & Salah	Multimodal personality trait analysis for explainable modeling of job interview decisions	2018
Kazim et al.	Systematizing audit in algorithmic recruitment	2021
Kim & Heo	Artificial intelligence video interviewing for employment: perspectives from applicants, companies, developer and academicians	2021
Kleinerman et al.	Supporting users in finding successful matches in reciprocal recommender systems	2021
Köhl et al.	Explainability as a non-functional requirement	2019
König & Langer	Machine learning in personnel selection	2022
Langer et al.	Spare me the details: How the type of information about automated interviews influences applicant reactions	2021
Langer & König	Explainability of artificial intelligence in human resources	2022
Lazzari et al.	Predicting and explaining employee turnover intention	2022
Lee	Applying Explainable Artificial Intelligence to Develop a Model for Predicting the Supply and Demand of Teachers by Region.	2021

Liem et al.	Psychology meets machine learning: Interdisciplinary perspectives on algorithmic job candidate screening	2018
Miller	"But why?" Understanding explainable artificial intelligence	2019
Ochmann & Laumer	Fairness as a determinant of AI adoption in recruiting: An interview-based study	2019
Ortega et al.	Symbolic AI for XAI: Evaluating LFIT inductive programming for fair and explainable automatic recruitment	2021
Park et al.	Designing fair AI in human resource management: Understanding tensions surrounding algorithmic evaluation and envisioning stakeholder-centered solutions	2022
Park et al.	Human-AI interaction in human resource management: Understanding why employees resist algorithmic evaluation at workplaces and how to mitigate burdens	2021
Pessach et al.	Employees recruitment: A prescriptive analytics approach via machine learning and mathematical programming	2020
Principi et al.	On the effect of observed subject biases in apparent personality analysis from audio-visual signals	2019
Robert et al.	Designing fair AI for managing employees in organizations: a review, critique, and design agenda	2020
Schumann et al.	We need fairness and explainability in algorithmic hiring	2020
Sekaran & Shanmugam	Interpreting the Factors of Employee Attrition using Explainable AI	2022
Singer & Cohen	An objective-based entropy approach for interpretable decision tree models in support of human resource management: The case of absenteeism at work	2020
Tambe et al.	Artificial intelligence in human resources management: Challenges and a path forward	2019
Tao et al.	Research on the Prediction of Employee Turnover Behavior and Its Interpretability	2021
Tippins et al.	Scientific, legal, and ethical concerns about AI-based personnel selection tools: a call to action	2021
Tsiakas & Murray-Rust	Using human-in-the-loop and explainable AI to envisage new future work practices	2022
Wang et al.	Personalized employee training course recommendation with career development awareness	2020
Wicaksana & Sukma	Human-explainable features for job candidate screening prediction	2017
Zhao et al.	Employee turnover prediction with machine learning: A reliable approach	2019

## Appendix B: Overview of the Measures Used in the Questionnaire

Measure (original name, items selected)	Used items	Source
<b>Choice effort</b> (Evaluation Costs, 2.-5. item)	1. I could not afford the time to fully evaluate relevant details.* 2. It was tough to compare the different employees.* 3. It was difficult for me to make this choice. 4. I concentrated a lot while making this choice.	Heitmann et al. (2007)
<b>Choice difficulty</b> (Choice Ease)	The task was ... 1. not at all difficult / extremely difficult. 2. not at all confusing / extremely confusing. 3. not at all overwhelming / extremely overwhelming.	Kelting et al. (2017)
<b>Attitude toward IS</b> (Information Value of the Web Site)	The information offered is ... 1. useful. 2. understandable. 3. sufficient.	Holzwarth et al. (2006)
<b>Uncertainty avoidance</b> (1., 3. item)	1. Security is an important concern in my life. 2. It is important to consider dissenting views when making personal and social decisions.	Erdem et al. (2006)
<b>Knowledge of AI/XAI</b> (Knowledge of the product class (Expert), 1., 2. item)	1. I understand the features of ___ enough to be considered an expert when evaluating different brands. 2. I know exactly what product characteristics are needed when buying a ___.	Alavi et al. (2016)
<b>Need for cognitive closure</b> (1., 3., 8., 42. item)	1. I think that having clear rules and order at work is essential for success. 2. I don't like situations that are uncertain. 3. I feel uncomfortable when I don't understand the reason why an event occurred in my life. 4. I dislike the routine aspects of my work (studies). (r)	Kruglanski et al. (1993)
(r) Reverse-scored. *Adapted from the original item.		