How Much AI Do You Require? Decision Factors for Adopting AI Technology

Completed Research Paper

Jonas Wanner
University of Würzburg
Würzburg, Germany
jonas.wanner@uni-wuerzburg.de

Kai Heinrich
Technische Universität Dresden
Dresden, Germany
kai.heinrich@tu-dresden.de

Christian Janiesch
University of Würzburg
Würzburg, Germany
christian.janiesch@uni-wuerzburg.de

Patrick Zschech
Technische Universität Dresden
Dresden, Germany
patrick.zschech@tu-dresden.de

Abstract

Artificial intelligence (AI) based on machine learning technology disrupts how knowledge is gained. Nevertheless, ML's improved accuracy of prediction comes at the cost of low traceability due to its black-box nature. The field of explainable AI tries to counter this. However, for practical use in IT projects, these two research streams offer only partial advice for AI adoption as the trade-off between accuracy and explainability has not been adequately discussed yet. Thus, we simulate a decision process by implementing three best practice AI-based decision support systems for a high-stake maintenance decision scenario and evaluate the decision and attitude factors using the Analytical Hierarchy Process (AHP) through an expert survey. The combined results indicate that system performance is still the most important factor and that implementation effort and explainability are relatively even factors. Further, we found that systems using similarity-based matching or direct modeling for remaining useful life estimation performed best.

Keywords: Artificial Intelligence, AI Adoption, Explainable AI, Analytical Hierarchy Process

Introduction

In 2017, Gartner declared 'Artificial Intelligence (AI) everywhere' a mega-trend and called it the most disruptive class of technologies over the next decade and its hype as well as its promise has not faded since (Panetta 2017). A major reason is today's ubiquitous use of IT and enhanced sensor technology that facilitate the generation of large amounts of data, providing an ideal starting point for knowledge discovery and improved decision support (Carvalho et al. 2019; Zschech et al. 2019). This new way of gaining knowledge holds great potential especially in critical areas such as cancer diagnostics in medical science, machinery and vehicle maintenance, or in autonomous driving.

To process the data and to gain new insights, advanced AI capabilities are used that are based primarily on algorithms from the field of machine learning (ML). Here, artificial neural networks (ANN) are particularly promising as they are able to identify complex nonlinear relationships within large datasets based on flexible mathematical models (Miller and Brown 2018). Especially in tasks related to high-dimensional data, such as classification and regression, ML algorithms show good applicability. By learning from previous computations and extracting regularities from given examples, they can produce accurate models and repeatable decisions. On the downside, AI-based decision support systems (AI DSS) have the disadvantage that the processing mechanisms and their results are often difficult to reproduce as ML

algorithms usually show black-box characteristics (Miller and Brown 2018). Thus, decision support based on a 'black box' AI model might not be fully explained or understood (Siau and Wang 2018). This could lead to users not accepting the system and consequently, not using it (Dam et al. 2018). Still, in domains where a lot of input information needs to be processed for decision making, experts have to rely on automated, intelligent decision support (Raouf et al. 2006). As a result, humans need to align their common cause and effect evaluation of the decision situation with the assessment reasoning of the system to accept the decision of the AI system (Miller 2019). This creates a dilemma between the demand for accurate vs. transparent DSS, which is the object of research on explainable AI (XAI) (Wanner et al. 2020). Consequently, XAI aims at a high degree of model explanation while maintaining the best performance possible (Turek 2016).

While research focuses on these two areas, practical issue such as environmental constraints and individual user preferences seem to be ignored. In particular, only cost and time seem to be considered as crucial in (IT) projects (Clancy 1995; Lech 2013). Hence, today, we observe rather isolated research on either model accuracy or model explainability. Yet in contrast, decisions towards adopting complex technologies are very complex and are subject to a range of cognitive bias, especially for applications that involve high-stake decisions and therefore are of high criticality (Das and Teng 1999). We understand high-stake decisions as those where a wrong or late decision can result in the loss of human lives.

With this research, we want to explore the human factor in AI adoption. We aim to understand the variables influencing the decision of choosing a particular AI technology for decision support in applications involving high-stake decisions (Stefani and Zschech 2018). Consequently, our research question is as follows:

RQ: *Which factors are of importance when choosing an AI DSS for high-stake decisions, and how can we characterize the trade-off between different AI DSS implementations?*

In order to answer this question, we derive *decision factors* and from the related domains ML, XAI, and IT project management. Then, we propose observable measurements for those factors. After modeling the decision as a hierarchy of decision factors and superior *decision categories*, we gather judgements on factor importance from a survey. Lastly, we simulate the decision process for the case of a high-stake maintenance scenario and weigh the proposed decision factors as a result of the survey.

Our research yields several contributions: First, we derive a theoretical construct of measurement variables from the interdisciplinary field of AI adoption, which is based on well-founded scientific theories of Information Systems and Business Research and can also form the basis for further AI research. Second, our research involves the derivation and implementation of best practices of AI DSS for the case of high-stake maintenance, which are highly relevant for practical application. Third, our end-user-based study results allow for a better understanding of the significance of decision criteria for AI adoption. On the one hand, this benefits research, which can focus on promising approaches in addition to further evaluation possibilities, and on the other hand, practice benefits from the decision model and proposed measurements for implementing their own decision process.

Our paper is structured accordingly: In the next section, we present the theoretical background and derive evaluation factors in terms of decision and attitude factors for AI-based system adoption. Subsequently, we depict our research methodology and describe the use case, followed by the introduction of a measurement model for the decision factors as well as the high-stake maintenance case. We then proceed to introduce the decision model, describe the survey setup, and present our results. In the last section, we summarize our contribution, discuss limitations, and present an outlook of further research opportunities.

Foundations, Related Work and Derivation of Evaluation Factors

AI-based Decision Support Systems

From an economic point of view, a decision is an intellectual effort to make the best possible choice to given circumstances. For a decision maker, information is the key to understand these circumstances and make the best possible judgement. Due to the increasing use of information systems in daily business, both the complexity and the amount of available information is growing (Bonczek et al. 2014). Research on decision support tries to address this by providing support for the decision-maker using modern techniques and optimization theory. To be more specific, a decision support system encapsulates complexity and allows

combining personal assessments about a necessary decision (problem) with computer output in a user-machine interface. In this way, it produces meaningful information for support in a decision-making process (Simonovic 1999). Through the years, these systems have developed into 'intelligent' systems called AI DSS (Bonczek et al. 2014). AI DSS are able to iteratively learn from empirical data, allowing them to find non-linear relations and complex patterns without being explicitly programmed (Bishop 2006).

Beyond the possibilities, however, there is the problem that decision makers must adopt and use the system and its underlying information for its advice to be effective. This leads to the research area of technology acceptance. A key construct of the related Unified Theory of Acceptance and Use of Technology (UTAUT) model is the user's performance expectancy. It is defined as the degree to which an individual perceives that using an information system will help him or her to attain an increase in his or her job performance (Venkatesh et al. 2003). AI research, as the backbone of AI DSS, is therefore concerned with the continuous improvement and expansion of algorithms. The basic quality of an AI algorithm is measured on two kinds of indicators: *prediction value (PV)* and *inference time (IFT)* (e.g., Lane et al. 2016). The former refers to the prognostic power of the algorithm model. The latter is defined as the trained model's delay in the calculation of a new data input in practice. Both can lead to consequential costs.

Thus, recent research efforts are particularly directed towards the development and application of ANN with increasingly deeper architectures often summarized as deep learning (DL) (Zschech et al. 2019). These show a high prediction value with parallel low inference time. Their multi-layered architecture allows them to be fed with high-dimensional raw input data and then automatically discover internal representations at different levels of abstraction (LeCun et al. 2015). DL approaches currently outperform traditional models for most applications, which constitutes them as state-of-the-art in the field of data-driven practical ML implementation, for example in the field of (smart) maintenance (Khan and Yairi 2018; Wang et al. 2018).

Demand of Explainable AI-Systems

The high performance of ML and especially DL models comes at certain costs, as their nested, non-linear structure creates a lack of transparency by constructing internal high-degree interactions between input features, which are difficult to disaggregate into a human-understandable form. In other words, it is not clear what information in the input data drives the models to generate their decisions. Therefore, they are typically regarded as 'black boxes' (Samek et al. 2017). Although not every practical task might require an explainable decision model, because it may be more important, for example, to prevent a failure than to understand its cause (Dragomir et al. 2009), the lack of explainability still poses problem for accepting an AI decision support system. For human intelligence, it is generally an important aspect to be able to explain the rationale behind one's decision, while simultaneously it can be considered as a prerequisite for establishing a trustworthy relationship (Samek et al. 2017).

Trust issues in the context of an agency theory setup can result from the information asymmetry between the AI DSS (agent) and its user (principal) (Kleine 2013). Due to the absence of knowledge about the inner functions of the AI DSS, the user cannot effectively validate whether the system operates within his interests or if is the subject of adversarial alteration (Heinrich et al. 2020). The missing observability of the decision process as a property of black-box models renders it difficult to detect possible deviations from the user's expectations. While a black-box AI DSS has no ability to 'act in its own interest' as it lacks an own agenda, its decision making process is not observable on its own, resulting in issues comparable to a moral hazard that can be avoided by increasing observability through measures of algorithmic self-signaling by the AI DSS. This call for improved observability of the inner decision logic of an AI DSS is in line with the XAI community's demand for explainable models while maintaining a high degree of model predictability (Turek 2016).

In this case, explainability is about the understanding of the reasoning behind a decision by a human observer (Dam et al. 2018). Thus, the degree of model explainability remains dependent on the respective end user. However, there is a distinction between two types of explanation: global explanation and local explanation (Dam et al. 2018). The former offers a *causal explanation (CE)*. It is about the process of making an AI-based decision traceable by providing transparency into the decision model, its components such as parameters, or its algorithm (Ras et al. 2018). For this purpose, the causal model of the algorithm is of great importance. Local explanation is about a post-hoc interpretability of an AI-based decision recommendation through textual explanations, graphical presentations, or examples (Lipton 2018). It aims at an appropriate type of *visual explanation (VE)* for intended users (Heinrich et al. 2019; Miller 2019).

(X)AI-based Projects in Practice

IT projects are said to fail particularly often (Lech 2013). In order to identify the failing causes and derive success factors of IT project management, there is a long-term study of the Standish Group called the 'CHAOS-study'. It was first published in 1994 and since then is continuously updated, with a total dataset of over 40.000 individual projects that have been scientifically investigated. The study distinguishes between successful, partially successful, and unsuccessful projects by measuring the compliance of three criteria: budget, time, and planned functionality (Clancy 1995). This is also confirmed by the empirical study of Lech (2013).

At the beginning of a project, at the stage of project planning, it is particularly important to define the envisaged framework of the three criteria. It has been shown that uncertainties, changing environmental factors, or dependencies result in non-compliance with budgets and schedules influence projects (Lech 2013). These should therefore be minimized as far as possible through a comprehensive understanding of the project intent and influencing factors. Regarding scientific theories, the technology acceptance theory and the theory of transaction costs are particularly important. The construct of user's effort expectancy of the UTAUT model of the technology acceptance theory addresses this. It is defined as the degree of ease associated with the use of an information system (Venkatesh et al. 2003). Referring to Ghalandari (2012), effort expectancy has a strong link between the effort, its achieved performance, and the rewards resulting. From a project management point of view, it is therefore particularly important during project planning to understand the effort required and the resulting performance benefits for a potential solution and its alternatives. In addition, considering the many extensions of the acceptance theory, we can identify two potential attitude factors that could alter a ranking of decision factors: *risk attitude* and technology anxiety, which in our case would be reflected by the *willingness to use an AI system* (Yang and Yoo 2004).

With regard to the future AI DSS, effort for the decision logic of the AI algorithm results from the *implementation time (IMT)* and the associated resource intensity, which depends on the *training time (TT)* of the AI model (Lane et al. 2016). Further, it is also important to understand the extent to which the project team has the *required skill level (RSL)* to be able to realize the potential implementation or if acquisition of external expertise is necessary. In general, the decision to outsource capabilities and resources is grounded, among other factors, in transaction cost theory (Alaghehband et al. 2011). More specifically, the RSL of an AI DSS acts as an effort indicator to be compared with the firm's capabilities and, therefore, can act as a proximity factor for costs.

We summarize the derived decision and attitude factors in Table 1 to complete *step (a)*.

Table 1: Evaluation Factors of AI DSS for High-stake Decisions

Decision Factors			
Research Domain	*Category*	*Factor*	*Kernel Theory*
Machine Learning	Performance	Prediction Value (PV)	Technology Acceptance Theory
		Inference Time (IFT)	
Explainable AI	Explainability	Causal Explanation (CE)	Agency Theory
		Visual Explanation (VE)	
IT Project Management	Effort	Implementation Time (IMT)	Technology Acceptance Theory; Transaction Cost Theory
		Training Time (TT)	
		Required Skill Level (RSL)	
Attitude Factors			
Risk Attitude		Willingness to Use AI	

Methodology and Use Case

Research Methodology

As outlined above, we focus on the selection problem of AI DSS motivated through three domains: IT project management, ML, and XAI. We approach the problem with an experiment based on three kernel theories: transaction cost theory, technology acceptance theory, and agency theory. Specifically, to gain insights into the user's decision rationale when choosing an AI DSS, we model the decision as a hierarchy of the decision

factors using the Analytical Hierarchy Process (AHP) method by Saaty (2008). The complete research methodology is depicted in Figure 1. We apply the method to compare *m* variations of the domain-specific AI DSS for high-stake maintenance.

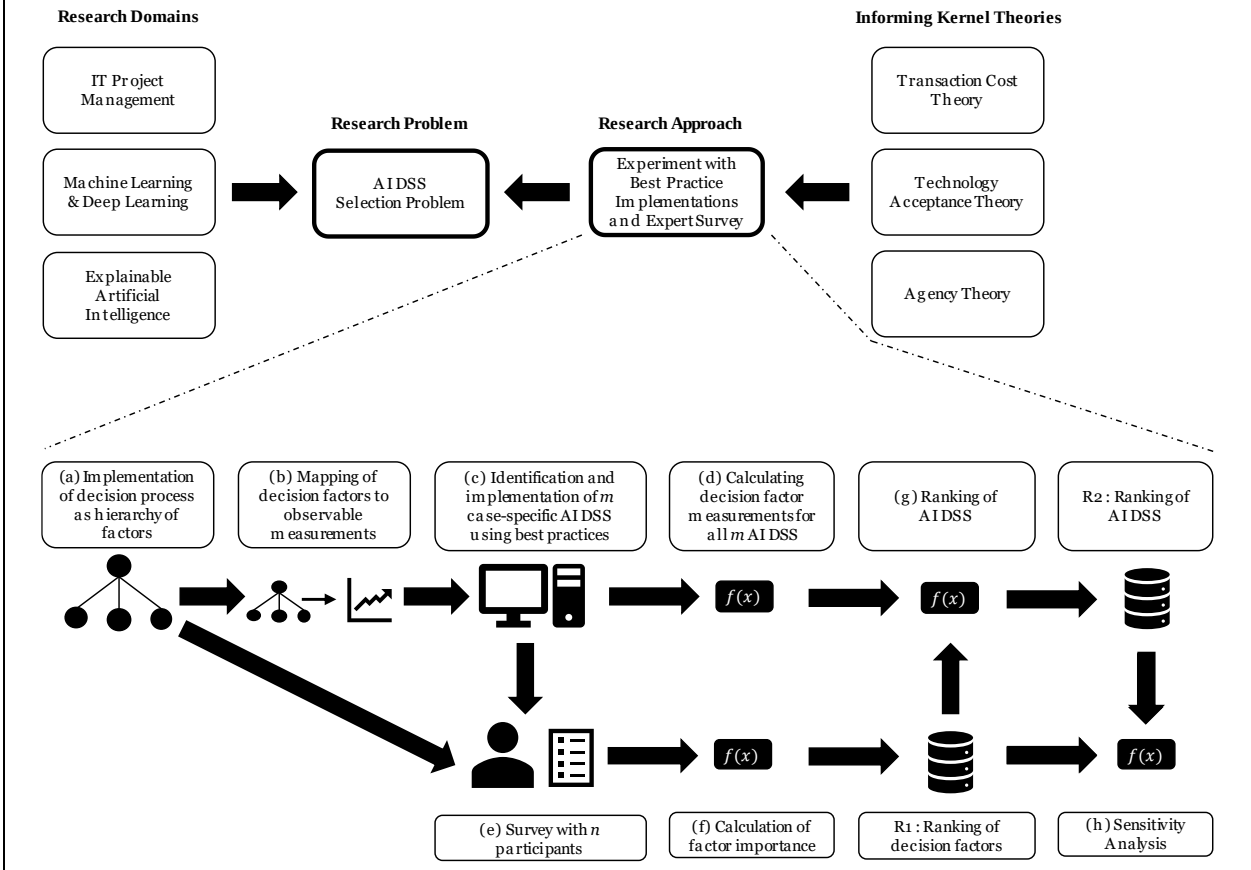

Figure 1. Research Methodology

Our general research approach using the AHP methodology is as follows: We (a) model the decision factors in the form of the hierarchy given in Table 1 (e.g., 'inference time' is a decision factor subordinate to the category 'performance'). Afterward, (b) the criteria on the lowest level are mapped to observable measures for a specific case (e.g., inference time is measured by the time it takes the algorithm to complete the classification of one element). After (c) implementing *m* variations of the AI DSS using best practice approaches, (d) the measures are calculated for every variation.

In parallel, we (e) use consistency-controlled pairwise comparisons between the decision factors from a survey to (f) obtain a transitive ranking of influence factors. In order to ensure the results are valid, we implemented a consistency check, as suggested by Saaty (2008), to ensure transitivity for every participant and his or her set of pairwise comparisons. The pairwise comparisons are carried out by presenting a questionnaire to *n* potential expert users of the AI DSS using the nine-point Saaty scale (1=equal importance, …, 9=extreme importance). As a result, we yield weights between 0 and 1 for the decision factors and the superior decision categories. In order to observe the moderating effects summarized in Table 1, we additionally included an item for each attitude factor using a five-point Likert scale.

Using the decision weight factors calculated in (f) and the measurements calculated in step (d), we can (g) calculate an overall rating of the *m* AI DSS. As a result of the AHP approach, we yield a first ranking (R1) of decision factor importance and, subsequently, a second ranking (R2) for the *m* AI DSS implementations. Finally, we conduct (h) a sensitivity analysis to explore the stability of the respective decisions and weights regarding the individual attitude variables. In addition, we point out that step (d) is independent of the use case so that the mapping can be adopted for other application areas that use AI DSS.

Use Case of High-stake Condition-based Maintenance

Industrial maintenance is intended to maintain and restore operational readiness of machinery to keep opportunity costs as low as possible. Related activities can be performed either i) after a break-down occurred, ii) in periodic cycles, or iii) in a condition-based manner. As unplanned downtime costs up to $250,000 per hour (Koochaki et al. 2011), or, in our specific case of high-stake maintenance cases, can even put human lives at risk (Raouf et al. 2006), a proactive condition-based maintenance (CBM) strategy seems to be the most useful approach (Kothamasu et al. 2006). For this purpose, comprehensive data collections are gathered and processed by a CBM system to assess the current state of the equipment and derive recommendations for the optimal time of intervention (Jardine et al. 2006; Veldman et al. 2011). In contrast to reactive strategies, divergent machine behavior can be detected and classified at an early stage by means of diagnostic techniques to avoid unnecessary work. Furthermore, by using suitable indicators and prognostic techniques, it is possible to determine a machine's future state or its remaining useful life (RUL), which is often referred to as predictive maintenance (Elattar et al. 2016; Peng et al. 2010).

To support CBM prognostic tasks, corresponding DSS can be based on different principles and approaches. This includes physical models, knowledge-based approaches, statistical methods, or AI-based approaches (Zschech 2018). Physical models are mathematical representations of physical processes, such as specific degradation laws. They can provide accurate health information, as the remaining quality of the component or system of interest, but they also require a thorough understanding of underlying mechanisms. Knowledge-based approaches are built on experiences from human experts. They can be formalized, for example, by means of domain-specific rules or heuristics, which, however, are difficult to obtain and hardly transferable to new contexts (Peng et al. 2010). Statistical methods, such as regression models, multivariate methods, and state-space models, use collected data observations to identify and model relationships for prognostic purposes. They are useful to assess risks quantitatively, but also heavily rely on assumptions of distributions, which in many real-world cases cannot be verified as fulfilled (Peng et al. 2010; Zschech et al. 2019). AI-based approaches cover a variety of ML techniques, such as support vector machines or different kinds of decision trees and ANN. They have the advantage of automatically exploiting hidden insights in vast amounts of observed data records (Elattar et al. 2016). Given the technological possibilities to collect large and multifaceted data assets in a simplified manner, AI-based approaches using ML can currently be considered as the most promising candidate for CBM decision support (Carvalho et al. 2019).

Measurement Model

In order to determine the overall decision, we need to map the decision factors to observable measures (in *step (b)*) and later calculate them for each of our m AI DSS (in *step (d)*).

Effort. Two characteristics can be quantified to approximate the effort of an algorithmic ML model: i) *implementation time* and ii) *training time*. The former calculates the time required to implement the selected algorithm, which is determined by the average of an estimation of several ML experts. This includes both preliminary steps, which we summarize as pre-processing, and the algorithmic realization, including re-engineering until a satisfactory solution is available. We define the implementation time (IMT), measured in person hours, by equation (1):

$$IMT = \sum_{i=1}^{n} (PP_i + AR_i) * \frac{1}{n} \quad (1)$$

averaging over all estimations n of the necessary time for pre-processing PP_i and the time needed to realize the algorithm AR_i itself over all experts i.

The algorithm's training time, on the other hand, indicates the resource intensity of a model for training. Training time (TT) is measured in seconds and determined by equation (2):

$$TT = \sum_{e=1}^{m} \sum_{b=1}^{n} |(tr_{ebt} - tr_{ebt+1}) + (te_{ebt} - te_{ebt+1})| \quad (2)$$

adding up the times needed to perform all training epochs e over all batches b as the number of samples given to the model before model's parameter updating, measured by the difference between the training's start time (tr_{ebt}) and end time (tr_{ebt+1}), and its related times needed to test the result ($te_{ebt} - te_{ebt+1}$).

A further characteristic that has a significant effect on the effort when implementing algorithms is the iii) required skillset (RSL). This cannot be quantified directly and, thus, needs to be substituted for

measurement. We use the required training as a proxy variable to do so and obtain the values based on the training courses on the platform *datacamp.com*. These courses were analyzed by data science experts that beforehand were familiarized with the prognostic approaches to filter out the courses of interest. Thus, we received a dictionary mapping the course short name and time to complete each of them (equation (3)):

(3) $RSL_c = \{\text{R|python basics: 8, ML basics: 8, ML regression: 4, DL: 12, Bayesian modeling: 8}\}$

We then accumulated the number of hours of all courses that would be necessary for a computer science student with no particular data science knowledge to implement the respective AI DSS, referred to as the choice c. We used the implementations and source code provided by the authors of the respective prognostic approaches identified in the next section[1].

Performance. Two characteristics quantify the performance of an algorithmic model: i) *prediction value* and ii) *inference time*. Prediction value (*PV*) refers to the prognostic power of an algorithm. We identified two measures, which are used in RUL data challenges: the '*PHM08' score* and the *root mean squared error* (RMSE) (Zschech et al. 2019). 'PHM08' prefers an early prediction to a late one. Thus, the scoring is asymmetric around the true time of failure, and late predictions are penalized more. The penalty itself expands exponentially with increasing error (Saxena et al. 2008). Equation (4) and (5) describe this:

(4) $$s = \sum_{i=1}^{N} s_i \qquad with \qquad S_i = \begin{cases} e^{\frac{d_i}{13}} - 1 \; for \; d_i < 0 \\ e^{\frac{d_i}{10}} - 1 \; for \; d_i \geq 0 \end{cases}$$

(5) $$d_i = \hat{L}_i - L_i^T$$

with s being the calculated score, N the number of units to be maintained and d_i the deviation of the estimated RUL $\hat{L}_i$ against the real RUL L_i^T. The preference for premature predictions corresponds to a risk-averse attitude and forces a high sensitivity to outliers in the exponential curve. A disadvantage is that the prognostic distance is not included in the scoring evaluation and algorithms are preferred that intentionally underestimate the RUL (Lim et al. 2014).

Thus, the RMSE is often used in parallel to measure the accuracy, punishing intentionally underestimated RUL, and in this way, it allows for a critical sub-evaluation. It is calculated using equation (6):

(6) $$RMSE = \sqrt{\sum_{i=1}^{n} \frac{(\hat{y}_i - y_i)^2}{n}}$$

adding all $(\hat{y}_i - y_i)^2$ squared differences of the real values against the calculated values in relation to the given sample size of n. The root obtains the actual deviation distance as the total error.

The inference time (*IFT*) describes the delay in the calculation of an algorithmic model in practical use. Thus, it is defined as the time a model takes between the input of new data and its calculated result, representing its real-time ability. It is measured in microseconds. *IFT* can be formulated using equation (7):

(7) $$IFT = \sum_{i=1}^{n} |t_i - t_{i+1}| * \frac{1}{n}$$

averaging over all time differences between the beginning of the calculation as the time of reading the new data t_i and the time the algorithm finished calculating the results t_{i+1}, which in turn represents the start time of the new calculation process.

Explainability. Both explainability characteristics i) *causal explanation (CE)* and ii) *visual explanation (VE)* can only be measured on an ordinal scale to create a ranking comparing the different solutions.

The former is about the ability of the algorithm to reveal the influence of each input variable with regard to the output. We measure this as a binary function, following the categorization of Yang et al. (2019). We set

[1] As a result, we found that every AI DSS implementation would require a basic set of courses in either R or Python (8 hours). Further, *similarity-based matching* (*AI DSS*$_3$) and *direct RUL* (*AI DSS*$_1$) require basic ML skills (4 hrs.), while *direct RUL* requires (*AI DSS*$_1$) at least two additional courses in DL (12 hrs.). *Indirect RUL* (*AI DSS*$_2$) requires Bayesian modeling (8 hrs.) in addition to basic ML course.

the measurement *causal explanation*=1 whenever a DL system can provide either a local and/or global intrinsic explanation manner, and we set *causal explanation*=0 when this is not the case.

Since the model visualization quality largely depends on the human's ability to interpret it, we chose to use the model visualizations of the *m* prognostic approaches (cf. next section) to our participants and let them rank the visualizations. Subsequently, we are calculating normalized weight scores for each AI DSS as a measurement for visual explanation.

Identification of AI DSS, Implementation, and Results

Identification of Best Practice AI DSS Using Datasets for Predictive Maintenance

In order to decide on the number and types of the *m* AI DSS for implementation, we need to identify best practices or at least promising examples of CBM using AI technology (in *step(c)*). As an AI-based CBM prognostic approach requires sufficiently extensive machinery data including failures to draw conclusions for early indicators, we decided to survey the most widely used datasets for high-stake maintenance decisions to identify the best practices of AI DSS.

We limited of search to predictive maintenance datasets that are publicly available, deal with high-stake maintenance cases, and have been used in a sufficient number of publications. We focused on the identification of existing comparative reviews first. We selected the following databases for our literature review: *IEEE Xplore Digital Library*, *EBSCOhost Academic Search Elite*, *EBSCOhost*, *Business Source Complete*, *ACM Digital Library*, *ScienceDirect*, and *Web of Science* using terminologies from predictive and condition-based maintenance combined with generic terms of datasets, comparisons, and reviews. Further, we have excluded terms from medical science, as these otherwise would have pre-dominated the results. As a result, we identified 25 datasets from five review articles. We used a hit popularity analysis to reveal the most popular datasets in research by the number of downloads and citations.

A closer analysis of the top three datasets highlight that *Turbofan (NASA)* and *PHM 2008* share the same data basis of NASA's Prognostics Data Repository (Elattar et al. 2016; Zschech et al. 2019): The so-called C-MAPSS data simulates large commercial engines (Richter 2012). It represents a high-stake decision case as a failure of engine turbines can lead to a crash of an airplane and related human casualties. To date, however, besides prognostic scoring, user acceptance has not been tested.

Using the taxonomy of Zschech et al. (2019), it is possible to make a distinction between three different data-based prognostic approaches for CBM using the C-MAPSS dataset. Among others, the taxonomy distinguishes the five existing scenarios of C-MAPSS. We consider this an essential limitation to ensure comparability and to identify best practices. For our purpose, we chose the dataset of the 'FD001' scenario as it is most commonly used in 60 out of 106 papers.

We differentiate between i) *direct RUL-mapping* ($AI\ DSS_1$), ii) *indirect RUL-mapping via health index (HI)* ($AI\ DSS_2$), and iii) *similarity-based matching* ($AI\ DSS_3$) (cf. Figure 2). In i) direct RUL-mapping, the training data is transformed into a multidimensional feature space first, using RULs to label the vectors. Then, a mapping between feature vectors and RUL is developed using methods of supervised ML (Ramasso and Saxena 2014). A total of k=21 publications of our subset chose this approach. From 2016, the majority apply different types of ANN with a tendency towards deeper architectures (Zschech et al. 2019). For ii) indirect RUL-mapping via HI, two mapping functions are combined. The first one maps sensor values to a HI for each training unit. The second links HI values to the RUL. Thus, a degradation model library is formed that serves as prior knowledge to update the parameters of the model corresponding to the new test instance (Ramasso and Saxena 2014). A total of k=28 publications of our subset chose this approach. The types of algorithms used differ. Thus, there is no clear favorite. For iii) similarity-based matching, historical degradation progressions (HI curves) are calculated and stored, forming a library with known failure times first. The RUL of a new instance is then estimated with the help of the most similar stored ones (Ramasso and Saxena 2014). A total of k=12 publications of our subset chose this approach. Various distance measures and geometric calculations are used to calculate the assessment of the similarity (Zschech et al. 2019).

In order to derive best practices for the three prognostic approaches, we looked at the performance ranking of the listed publications. Accuracy-based metrics have been used especially in the related data challenges to compare the different algorithm coding. Ordered by the 'PHM08' score, the two highest-scoring

publications belong to the category of similarity-based matching with another similarity-based matching paper in the top ten. All other seven top ten publications use direct RUL-mapping. An algorithmic analysis of all publications revealed the technical best practices per prognostic approach: For the i) direct RUL-mapping, using a long-short-term memory network (LSTM) or a convolutional neural network (CNN) seems promising. This is in accordance with four out of the top ten publications by 'PHM08' score, where three are based on an LSTM. Focusing on ii) indirect RUL-mapping via HI, using an extrapolation is a good choice. For iii) similarity-based matching, an LSTM encoder-decoder combined with a curve matching and a similarity-based reasoning with a polygon clipping is favorable. This is in line with the two best-performing contributions (Malhotra et al. 2016; Ramasso and Saxena 2014).

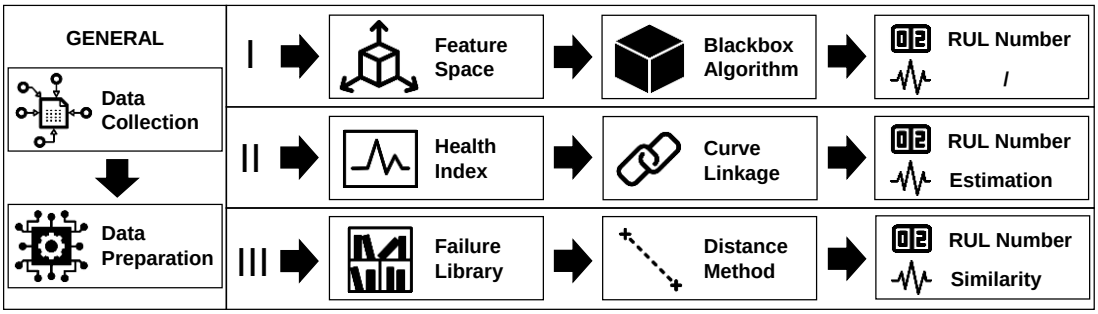

Figure 2. Model Explanation per Prognostic Approach

Implementation of Best Practice AI DSS

We have based our own prototypical implementations (in *step (c)*) on these best practices of prognostic approaches. If their documentation was not available in a comprehensible way, we have tried to follow a comparable study. It was particularly important for us to understand the technical challenges that the implementation entails and the opportunities that arise as a result. For the i) direct RUL-mapping approach, we have implemented an LSTM. For ii) indirect RUL-mapping via HI, we have used an algorithm based on extrapolation. The iii) similarity-based matching was reconstructed using a squared polynomial curve fitting. These three approaches represent our m=3 AI DSS for the case of high-stake maintenance.

Direct RUL-mapping. For the RUL target function, we determined the RUL value by the difference between the total runtime of the unit and the individual cycle numbers. This corresponds to a decreasing trend. We set the maximum RUL estimation to 125, in accordance with Zhu et al. (2018) and Malhotra et al. (2016). The size of the time window is defined by N_{tw}=30 (Li et al. 2018; Zhu et al. 2018). This is equivalent to the window size in Zhu et al. (2018) and has been proven by the results of Li et al. (2018), who made a comparison of time window sizes for the C-MAPSS dataset. We used the *mean squared error* (MSE) as the loss function and selected all 21 sensors as inputs. The architecture of our artificial network was based on Zheng et al. (2017) as the best ranked LSTM solution according to the 'PHM08' score. We have implemented a total of five layers: two LSTM layers, followed by two full layers and another final full layer. For the parameterization of the LSTM layers, own cross-validation studies have shown that two 64-units layers achieve more robust results than the 64- and 32-unit version of the one in Zheng et al. (2017). The full layers were set to 8-units each. Further, Zheng et al. (2017) mentioned the usage of a dropout rate without specifying its value. Our test set shown the best result for *dropout*=0.2 and a final *dropout*=1.

Indirect RUL-mapping via HI. Our method for the indirect RUL-mapping via HI is based on the work of Coble and Hines (2011). They have proposed a dynamic Bayesian updating procedure that allows a priori information from the training data to be incorporated into the extrapolation procedure of the test data to obtain the model parameters. So, the HI is calculated with linear regression, using the first 10% of training instances as 1 and the last 10% as 0. Comparing different time windows of preliminaries, we settled with N_{tw}=15 (Li et al. 2018). We smoothed the degradation curve using locally weighted scatterplot smoothing (LOWESS). It assigns the highest weight to the value to be smoothed and lower weights to more distant values. To forecast the RUL, the degradation curve is plotted up to a predetermined limit value indicating the extrapolated error status. We set this limit value to h=0. For the necessary curve fitting, we used dynamic Bayesian updating with a polynomial of the 2nd order. Coble and Hines (2011) proofed that while

an exponential polynomial would be physically more useful, a quadratic polynomial is more robust in terms of noise and results in a better fit.

Similarity-based matching. The implementation of our similarity-based algorithm follows the work of Malhotra et al. (2016), who achieved the second-best score for the dataset 'FD001'. First, we built a reference library. Thus, we created one-dimensional degradation curves in the form of HIs from the multidimensional training data. Our implementation of the first similarity-based approach was geared to Wang et al. (2017). We calculated the new HI using linear regression and a squared polynomial for curve matching. The parameters are determined by *least-square fitting*. We calculated the similarity score using *information fusion* and considering the time lag between training and test data. The weighting parameters μ and θ are determined as μ=0.8 and θ=0.2 (Wang et al. 2017). The maximum RUL estimation is set to 125. Our second implementation was geared to Malhotra et al. (2016). The HI is calculated by linear regression. Here, no curve matching is applied to assign the HI to similar functions of the library. Instead, we smoothed the historical HI degradation progressions using the LOWESS method with a time window of N_{tw}=15 (Li et al. 2018). In addition, we considered the time lag. We set the maximum RUL estimation to 125 (Malhotra et al. 2016; Zhu et al. 2018). We calculated the similarity score using Euclidean distance and set the scaling similarity measure to λ=0.0005. We determined the weighting parameters as μ=0.8 and θ=0.2 (Wang et al. 2017).

Results. The quantifiable measures of each prototypical AI DSS solution can be found in Table 2 (completing *step (d)*). All algorithms have been deployed ten times, taking their average values. As hardware backbone, we used an Intel Core i7-6700HQ CPU (4x 2.60 GHz), with an NVIDIA GeForce GTX 960M and 16 GB RAM, operating under Windows 8.1 pro.

Table 2: Measurement Results for High-stake Maintenance AI DSS

Prognostic Approach (AI DSS)	*Effort*			*Performance*			*Explainability*	
	IMT	TT	RSL	PHM08	RMSE	IT	CE	VE
Direct RUL	8 hrs	313.9 sec	20 hrs	270	13.51	1562 µsec	No	0.10
Indirect RUL	15 hrs	44.9 sec	16 hrs	775	21.23	6006 µsec	Yes	0.52
Similarity-based	25 hrs	57.2 sec	12 hrs	409	14.17	5345 µsec	Yes	0.38

Decision Model and Results Discussion

Decision Model

Following *step (a),* we first build a hierarchical model of the decision process, where the choices are the m=3 AI DSS and the decision factors of Table 1. The hierarchy of factors is depicted in Figure 3. The decision factors are depicted by white filled boxes while the actual choices of AI DSS implementations are depicted as black filled boxes.

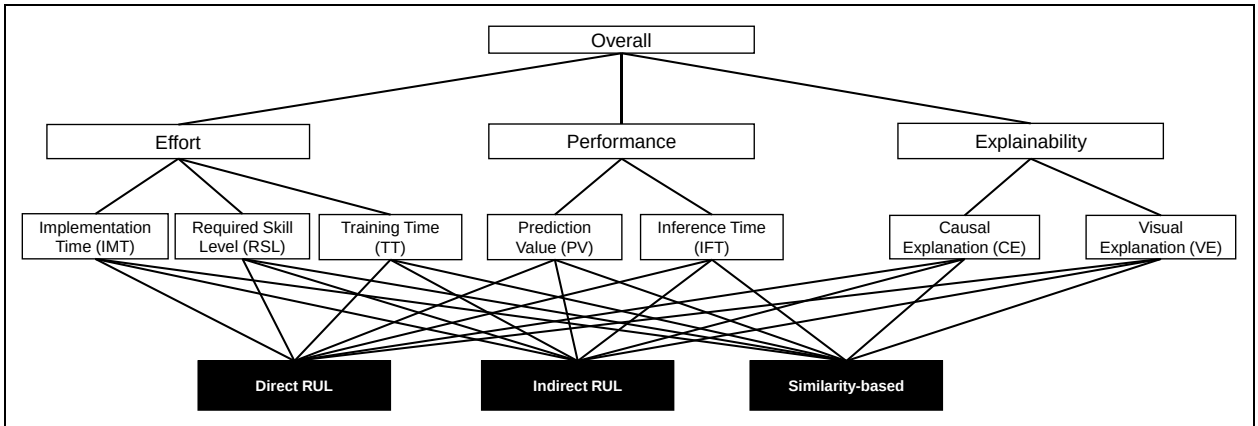

Figure 3. Factor Hierarchy for AHP

The questionnaire (in *step (e)*) started with items regarding basic demographic information about gender, age, region, industry sector experience, company size, current position, and work experience. It also included views of the two moderating attitude factors (willingness to take risks and AI at the workplace). Then, we introduced the use case of establishing preferences for a future AI DSS at the workplace for high-stake maintenance of airplane turbines. In the second part of the survey, we first asked the participants to rank visual explanations of the m=3 AI DSS. Examples of the approaches are depicted in Figure 4.

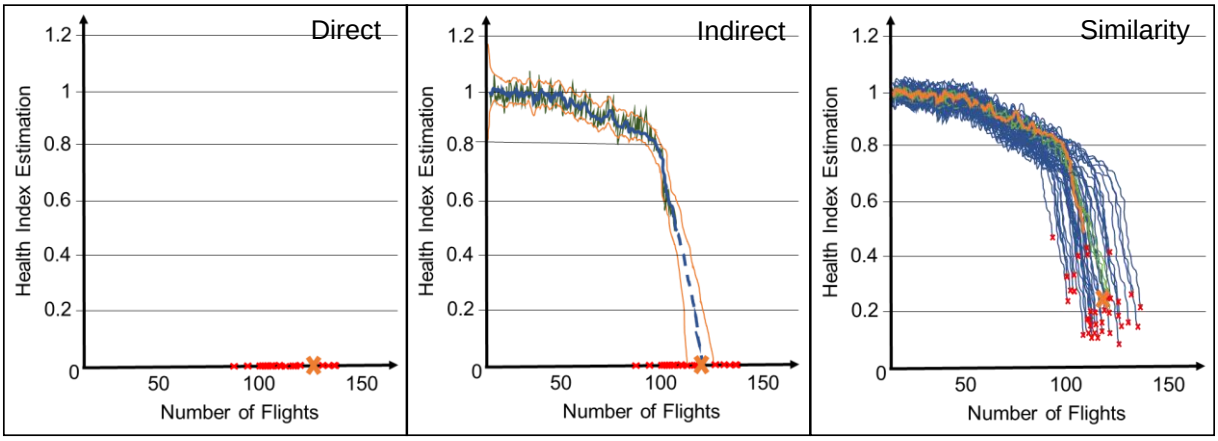

Figure 4. Visual Explanation of Each Prognostic Approach (AI DSS)[2]

Second, we asked for pairwise comparisons of the decision factors, and third, we asked for pairwise comparisons of the categories (for both cf. Table 1). For example, we asked whether participants considered the inference time or the prediction value of an AI DSS to be more important and, if so, by how much on the aforementioned nine-point Likert scale. In doing so, we created a weighted hierarchy of decision factors as well as a weighted hierarchy of the summative categories.

Then, we used the platform *Prolific.co* to recruit n=165 participants[3]. The platform enabled us to screen for professionals in the fields of manufacturing, aviation, and other related industrial fields. We cross-checked this in the survey using demographic items.

After receiving the initial data, we checked all questionnaires by applying the following inclusion criteria: (i) The participants needed at least 5 minutes to complete the survey (with 5 minutes being the minimum time determined to understand all given information), (ii) the participants had at least 2 years of experience in the field of maintenance, and (iii) the answers were not trivial due to primacy or recency bias or simply carelessness. After applying criteria (i) (-45), criteria (ii) (-47) and criteria (iii) (-0), we ended up with a final sample of n=73 (male=44, female=29; with age of $\bar{x}$=41, M=36, SD=11). 67% of the participants originated from Europe, 30% came from North America, and only 3% from other continents.

Our verification for experience in industrial maintenance showed a high experience in the topic of interest ($\bar{x}$=10, M=8, SD=6). The duration to complete the survey was in min. $\bar{x}$=9:28, M=7:54, SD=6:13. We additionally implemented a consistency check using Saaty's proposed method by calculating the consistency index and removing all judgments with an unacceptable inconsistency rate of >10%. No such inconsistencies were found, which can be explained by the straightforward design of only five and three pairwise comparisons, respectively.

We accumulated the n judgments based on the geometric mean approach described by (Dijkstra 2013). Multiplying the weights from R1 with the respective normalized measurements yields the overall rankings

[2] Fig. 3 shows an example for the visual presentation of results of the AI DSS implementations. The y-axis represents the health status estimated by the AI system. The x-axis indicates the amount of flights of the turbine. Red crosses represent the failures known to the system from the training data, whereas the large orange cross is the calculated failure of the turbine, which implies its RUL.

[3] The financial incentive for complete surveys was approx. $14/hr.

R2. Since we have two indicators of prediction value, we use the average of the RMSE and PHM08 as the AHP input for prediction value. We summarize the results in Table 3[4]. In addition, we provide the AHP weights for high- and low-risk attitude as well as high and low willingness to accept AI at the workplace to enable a sensitivity analysis regarding the attitude factors. For the *low* group of the attitude factors, only values 1 and 2 of the Likert scale were selected, while for the *high* group, only values 4 and 5 were selected.

Discussion of Results R1 and R2

R1. Overall (i.e., in *step (g)*), we found that *performance* is the factor that was selected to be most important for the decision which AI DSS to choose. Although this result was expected for a case of high-stake maintenance, the magnitude of the decision weight was lower than expected in comparison with literature that almost entirely suggests that measures of prediction value (however they might be expressed) along with inference time are the only factors used in evaluation (La Cava et al. 2019).

Table 3: Results of the AHP, Including Attitude Effects

		All	*Risk Attitude*		*Willingness to Use AI*	
Category	*Factor*	(n=73)	Low (n=25)	High (n=19)	Low (n=7)	High (n=38)
Decision Factors (R1)						
Effort	IMT	0.22	0.21	0.21	0.36	0.24
	RSL	**0.55**	**0.58**	**0.51**	**0.41**	**0.53**
	TT	0.23	0.21	0.28	0.22	0.23
Performance	PV	**0.87**	**0.92**	**0.70**	**0.84**	**0.90**
	IFT	0.13	0.08	0.30	0.16	0.10
Explainability	CE	0.25	0.37	0.27	0.39	0.21
	VE	**0.75**	**0.63**	**0.73**	**0.61**	**0.79**
Overall (Factors)	Effort	0.19	0.20	0.24	0.27	0.17
	Performance	**0.61**	**0.63**	**0.53**	**0.52**	**0.59**
	Explainability	0.20	0.17	0.23	0.21	0.22
AI DSS Decision (R2)						
Overall (Alternatives)	Direct RUL	0.35	0.35	0.34	0.35	0.34
	Indirect RUL	0.30	0.29	0.31	0.31	0.30
	Similarity	0.35	0.36	0.34	0.35	0.35

A detailed look at the performance category shows that measures for system prediction quality (e.g., RMSE) is the main factor. This result holds for low-risk attitude and both willingness attitudes (low and high). However, we discovered a significant change in decision weights when high-risk attitude participants decided which system to use. It shifts from 0.92 (low risk) to only 0.70, giving much more weight to the inference time. We expect this result, as sometimes risk-takers will ignore the probability (risk) that the AI system is wrong to a higher degree than risk-averse persons. Risk-takers gamble for a higher reward and for a situation with high performance, expecting the system still to put out a good-enough value for the maintenance professional to make the right decision at lower inference times and therefore maximizing utility. We see the opposite for risk-averse people, where inference time does not matter at all (only 8% decision weight) compared with the system being able to predict a value close to the true RUL. As mentioned before, we found that *effort* and *explainability* are equally important. While this is surprising from a scientific point of view, it was expected from a practitioner's viewpoint. The theory on explainability of AI DSS and, thus, the need to implement additional XAI algorithms or chose transparent models from the outset while partially ignoring performance, is an isolated viewpoint as is the theory on performance of AI

[4] Bold numbers indicate largest decision weights per category. We provide details on our data and calculations as a digital supplement at https://doi.org/10.13140/RG.2.2.18577.45928.

DSS (Rudin 2019). Only very few scientific articles are concerned with the actual cost of AI DSS or the comparison to alternatives (Elliott and Griffiths 1990; González-Rivero et al. 2020; Partel et al. 2019). While research for models to determine the value of an information system is a core IS research task, value, and costs estimates for AI DSS are still scarce (Matlin and Land O'Lakes 1979). This is partially explained by the fact that AI DSS typically introduce some degree of automation within a corporate information system landscape, and therefore it is hard to capture all long term benefits and cost savings. In addition, the recent AI and XAI literature is rarely concerned with value and costs of such systems as a core topic. However, our results reflect that in practice, as expected, this factor is attributed with some importance. Specifically, the required skill level with a decision weight of 0.58 stands out since it is seen as the biggest cost driver apart from implementation and training time, which were found to be approximately equal at 0.20. This can be explained by the fact that these criteria are not necessarily distinguishable by maintenance professionals who are not familiar with the concept of developing and training an AI DSS. We can detect a decision weight increase regarding the factor implementation time for the case of people with a low willingness to use AI. This could be associated with the negative expectation of particularly long implementation times for AI DSS. Among many reasons, this could be caused by own experience with a bad state of facilitating conditions, which, according to Venkatesh et al. (2003), influences the acceptance of the technology.

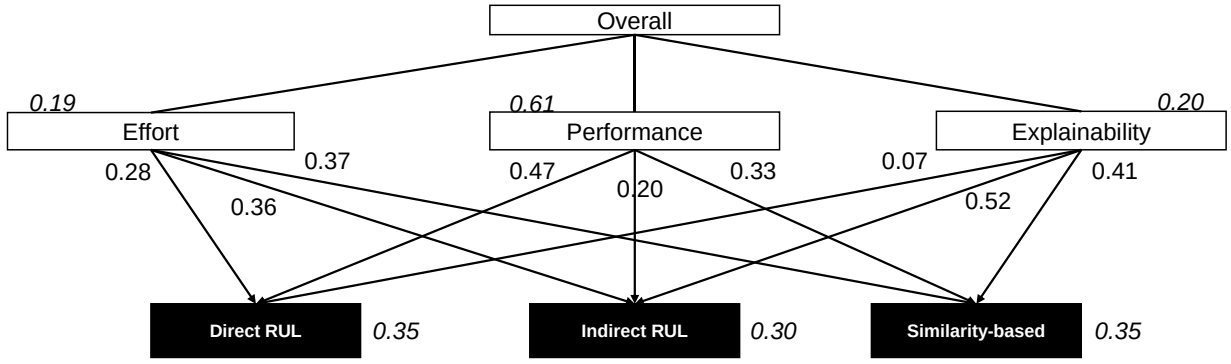

Figure 5. Decision Weights for the AI DSS Decision (all, *n*=73)

Within the category of *explainability,* we can distinguish between the ability of a model to connect its results to the input variables and therefore the ability to create transparency as to what caused the decision (causal explanation) and the ability to justify the 'correctness' of the results visually incorporating domain-specific knowledge already known to the user due to his or her experience. Surprisingly, the justification of the results seems to be much more important than the causal explanation for the system's output. This could be since the sheer number of sensor input variables specific to our case would overwhelm even expert users, while the visualization gives a clear degradation trajectory that is understood by all maintenance experts. In addition, the indication of variable importance to choose a value A over B expressed by a statistical measure is not the same as identifying a real root cause, which the AI DSS is usually not able to determine since it cannot identify the natural context of the problem and is not expected to do so by its users (Chaczko et al. 2020). We also find that for people with low willingness to use AI, the causal explanation and, therefore, the transparency of the system regarding its decision process is more important than for the other groups. This can be attributed to the fact that among other reasons, those people could have trust issues motivated by either previous experience or biased media exposure. Understanding how the system calculates the RUL and comparing it with their own experience and 'way of doing it', could, therefore, improve this relation and, thus, it is deemed more important by that group (Miller 2019).

R2. We depict the results of the decision process (R2) in the lower part of Table 3 and show that across all attitude variations, direct RUL and similarity-based matching are tied with scores of approx. 0.35 and indirect RUL is inferior with a weight of 0.30.

The choice weights of the overall factors that lead to those final scores R2 are depicted in Figure 5 for all participants (*n*=73). It is clearly shown that the *direct RUL* achieves its score purely by the performance advantage (0.47). This is in accordance with the fact that state-of-the-art black-box models based on DL

achieve high task-based prediction values. Improving the explainability of the model with additional XAI algorithms to justify the results and assign input variables weights towards the output would further strengthen this approach. The *similarity-based matching* only achieves mediocre performance but scores high in explainability (0.41) and therefore is tied with direct RUL at a normalized decision score of 0.35. In comparison to the direct RUL, which uses a DL model, the similarity-based approach takes about three times as long to implement than the direct RUL while requiring a lower skill level. The low implementation time of the direct RUL DL model can be attributed to the fact that DL has been gaining a lot of attention from non-AI professionals resulting in a considerable boost of high-level programming packages that can assist in building DL models more quickly. In contrast, the enormous training time of the direct RUL approach and the rather advanced required skillset explains its low overall score in effort (0.28) and, thus, direct RUL is inferior to the other approaches in terms of effort. Although the *indirect RUL* scores highest in explainability, the overall rather low importance of these factors, according to the survey participants, the low performance and the rather high effort renders it inferior to the other two models. Our results clearly show that the often-proposed trade-off between explainability and performance, while existent, is heavily skewed towards performance in the case of high-stake decisions.

Summary and Outlook

In this paper, we simulated a decision process and implemented an AHP to gain insights into the weights of decision factors for choosing an AI DSS. Therefore, we first derived decision factors from the three domains of ML, XAI, and project management to reflect all aspects of AI DSS applications in the context of a scenario with high-stake decisions where human lives are at risk. Second, we proposed a measurement model for those decision factors. Third, we implemented the decision process for a case of high-stake maintenance, more specifically, a maintenance case in aviation with the goal of preventing airplane turbine failure (called C-MAPSS). Fourth, we conducted a survey to understand the weights attributed to the decision factors and integrated this with measurement results from our implemented AI DSS.

We found that performance is still the overarching factor with effort and explainability tied in second place. In addition, we found that attitude factors influence the importance of decision factors, especially when people show low willingness to adopt AI and therefore have predispositions. For example, risk-takers valued inference time at the cost of prediction value more significantly and risk-averse experts, as well as experts with a low willingness to use AI, valued causal explanations to understand the AI model at the cost of visual explanations of decisions. Further, we found that direct RUL-mapping and similarity-based matching perform best with direct RUL-mapping benefiting from its prediction value and similarity-based matching from its ability to explain decisions.

With this research, we provide multiple theoretical and practical contributions. First, we provide well-founded and operationalized measurement variables for AI adoption that can form the basis for further AI research. Second, we have identified best practices of AI DSS for high-stake maintenance decisions, which yield high practical value. Third, our expert study enables an understanding of the significance of decision factors and attitude factors for choosing an AI DSS or, in general, for AI adoption. Fourth, we were able to draw conclusions, which factors should be focused on when certain risk attitudes and willingness or resistance to use AI can be assumed. This benefits research as scientists can use our results to better understand the trade-off between effort, performance, and explainability and devise further holistic evaluations. Likewise, practice benefits from the decision model and measures when implementing their own decision process.

As with all experimental research, we face some limitations. First, this is merely a descriptive study regarding the case of high-stake maintenance. Although the criticality was emphasized in the study, we cannot guarantee that it was always fully incorporated into the participant's decisions, as it would have if they were faced with a real-world situation. Second, we kept the description straightforward using a general predisposition towards AI systems and avoided adding detailed attitude factors such as fear of AI or AI-independent factors that can be derived from the UTAUT literature such as facilitating conditions or previous experience alongside hedonic motivation or habit. Third, we were aiming to avoid questionnaire bias from unintended questions regarding the case. Therefore, we set the criticality and the environmental constraints as static factors within the case. However, this removed the ability to implement a group design of experiment to further investigate special trade-offs when criticality changes or other environmental constraints are present.

While we applied the decision process to a specific case to demonstrate it, it can be used for the general problem of choosing an AI DSS for a case with high-stake decisions. The results of this study can be used in both normative and prescriptive decision analysis regarding AI systems as a starting point to investigate the trade-offs of decision factors for specific domains or even domain-independent contexts. Thus, for further generalization, we plan to extend our study to additional settings, such as medical diagnostics and autonomous driving. As such, we want to examine whether our results also apply to similar high-stake decision scenarios. Likewise, we will investigate uncritical environments as a counterpart to isolate the effect of a decision scenario's criticality. For this purpose, our study design provides a sufficiently generic basis to be applied to any conceivable domain and decision context. Furthermore, we plan to extend the study design to include multiple stakeholder groups within an organization by using the methodology of the Analytical Networking Process (ANP), which is an extension of the AHP that allows for consideration of decision clusters (e.g., allowing different stakeholders).

In addition, our study also raises awareness of the preferences of practitioners when it comes to obtaining an AI system. With the large focus on performance, specific issues like AI security and diagnosing failure in critical applications through explainability might need to be communicated more thoroughly through means of IS strategy (e.g., through corporate constraints) to be more resilient when facing these new challenges in the age of AI.

Acknowledgement

This research and development project is funded by the Bayerische Staatsministerium für Wirtschaft, Landesentwicklung und Energie (StMWi) within the framework concept "Informations- und Kommunikationstechnik" (grant no. DIK0143/02) and managed by the project management agency VDI+VDE Innovation + Technik GmbH.

References

Alaghehband, F. K., Rivard, S., Wu, S., and Goyette, S. 2011. "An Assessment of the Use of Transaction Cost Theory in Information Technology Outsourcing," *The Journal of Strategic Information Systems* (20:2), pp. 125-138.

Bishop, C. M. 2006. *Pattern Recognition and Machine Learning*. New York: Springer.

Bonczek, R. H., Holsapple, C. W., and Whinston, A. B. 2014. *Foundations of Decision Support Systems*. London: Academic Press.

Carvalho, T. P., Soares, F. A., Vita, R., Francisco, R. d. P., Basto, J. P., and Alcalá, S. G. 2019. "A Systematic Literature Review of Machine Learning Methods Applied to Predictive Maintenance," *Computers & Industrial Engineering* (137), pp. 106024-106034.

Chaczko, Z., Kulbacki, M., Gudzbeler, G., Alsawwaf, M., Thai-Chyzhykau, I., and Wajs-Chaczko, P. 2020. "Exploration of Explainable Ai in Context of Human-Machine Interface for the Assistive Driving System," *Asian Conference on Intelligent Information and Database Systems*, Cham, pp. 507-516.

Clancy, T. 1995. "CHAOS," in: *The Standish Group Report*.

Coble, J., and Hines, J. W. 2011. "Applying the General Path Model to Estimation of Remaining Useful Life," *International Journal of Prognostics and Health Management* (2:1), pp. 71-82.

Dam, H. K., Tran, T., and Ghose, A. 2018. "Explainable Software Analytics," *International Conference on Software Engineering: New Ideas and Emerging Results*, Gothenburg, pp. 53-56.

Das, T., and Teng, B. S. 1999. "Cognitive Biases and Strategic Decision Processes: An Integrative Perspective," *Journal of Management Studies* (36:6), pp. 757-778.

Dijkstra, T. K. 2013. "On the Extraction of Weights from Pairwise Comparison Matrices," *Central European Journal of Operations Research* (21:1), pp. 103-123.

Dragomir, O. E., Gouriveau, R., Dragomir, F., Minca, E., and Zerhouni, N. 2009. "Review of Prognostic Problem in Condition-Based Maintenance," *European Control Conference (ECC)*, Budapest, pp. 1587-1592.

Elattar, H. M., Elminir, H. K., and Riad, A. 2016. "Prognostics: A Literature Review," *Complex & Intelligent Systems* (2:2), pp. 125-154.

Elliott, D., and Griffiths, B. 1990. "A Low Cost Artificial Intelligence Vision System for Piece Part Recognition and Orientation," *International Journal of Production Research* (28:6), pp. 1111-1121.

Ghalandari, K. 2012. "The Effect of Performance Expectancy, Effort Expectancy, Social Influence and Facilitating Conditions on Acceptance of E-Banking Services in Iran: The Moderating Role of Age and Gender," *Middle-East Journal of Scientific Research* (12:6), pp. 801-807.

González-Rivero, M., Beijbom, O., Rodriguez-Ramirez, A., Bryant, D. E., Ganase, A., Gonzalez-Marrero, Y., Herrera-Reveles, A., Kennedy, E. V., Kim, C. J., and Lopez-Marcano, S. 2020. "Monitoring of Coral Reefs Using Artificial Intelligence: A Feasible and Cost-Effective Approach," *Remote Sensing* (12:3), pp. 489-510.

Heinrich, K., Graf, J., Chen, J., Laurisch, J., and Zschech, P. 2020. "Fool Me Once, Shame on You, Fool Me Twice, Shame on Me: A Taxonomy of Attack and De-Fense Patterns for AI Security," *European Conference on Information Systems (ECIS)*, AIS Virtual Conference Series, pp. 166.

Heinrich, K., Zschech, P., Skouti, T., Griebenow, J., and Riechert, S. 2019. "Demystifying the Black Box: A Classification Scheme for Interpretation and Visualization of Deep Intelligent Systems," *Americas Conference on Information Systems (AMCIS)*, Cancún, pp. 1-10

Jardine, A. K., Lin, D., and Banjevic, D. 2006. "A Review on Machinery Diagnostics and Prognostics Implementing Condition-Based Maintenance," *Mechanical Systems and Signal Processing* (20:7), pp. 1483-1510.

Khan, S., and Yairi, T. 2018. "A Review on the Application of Deep Learning in System Health Management," *Mechanical Systems and Signal Processing* (107), pp. 241-265.

Kleine, A. 2013. *Entscheidungstheoretische Aspekte Der Principal-Agent-Theorie*. Heidelberg, Germany: Springer-Verlag.

Koochaki, J., Bokhorst, J., Wortmann, H., and Klingenberg, W. 2011. "Evaluating Condition Based Maintenance Effectiveness for Two Processes in Series," *Journal of Quality in Maintenance Engineering* (17:4), pp. 398-414.

Kothamasu, R., Huang, S. H., and VerDuin, W. H. 2006. "System Health Monitoring and Prognostics—a Review of Current Paradigms and Practices," *The International Journal of Advanced Manufacturing Technology* (28:9-10), pp. 1012-1024.

La Cava, W., Williams, H., Fu, W., and Moore, J. H. 2019. "Evaluating Recommender Systems for Ai-Driven Data Science," *arXiv preprint arXiv:1905.09205*.

Lane, N. D., Bhattacharya, S., Georgiev, P., Forlivesi, C., Jiao, L., Qendro, L., and Kawsar, F. 2016. "Deepx: A Software Accelerator for Low-Power Deep Learning Inference on Mobile Devices," *International Conference on Information Processing in Sensor Networks (IPSN)*, Vienna, pp. 1-12.

Lech, P. 2013. "Time, Budget, and Functionality? - It Project Success Criteria Revised," *Information Systems Management* (30:3), pp. 263-275.

LeCun, Y., Bengio, Y., and Hinton, G. 2015. "Deep Learning," *Nature* (521:7553), pp. 436-444.

Li, N., Lei, Y., Yan, T., Li, N., and Han, T. 2018. "A Wiener-Process-Model-Based Method for Remaining Useful Life Prediction Considering Unit-to-Unit Variability," *Transactions on Industrial Electronics* (66:3), pp. 2092-2101.

Lim, P., Goh, C. K., Tan, K. C., and Dutta, P. 2014. "Estimation of Remaining Useful Life Based on Switching Kalman Filter Neural Network Ensemble," *Annual Conference of the Prognostics and Health Management Society*, Fort Worth, pp. 2–9.

Lipton, Z. C. 2018. "The Mythos of Model Interpretability," *Queue* (16:3), pp. 31-57.

Malhotra, P., TV, V., Ramakrishnan, A., Anand, G., Vig, L., Agarwal, P., and Shroff, G. 2016. "Multi-Sensor Prognostics Using an Unsupervised Health Index Based on Lstm Encoder-Decoder," *arXiv preprint arXiv:1608.06154*.

Matlin, G., and Land O'Lakes, I. 1979. "What Is the Value of Investment in Information Systems?," *MIS Quarterly* (3:3), pp. 5-34.

Miller, D. D., and Brown, E. W. 2018. "Artificial Intelligence in Medical Practice: The Question to the Answer?," *The American Journal of Medicine* (131:2), pp. 129-133.

Miller, T. 2019. "Explanation in Artificial Intelligence: Insights from the Social Sciences," *Artificial Intelligence* (267), pp. 1-38.

Panetta, K. 2017. "Enterprises Should Explain the Business Potential of Blockchain, Artificial Intelligence and Augmented Reality." *Top Trends in the Gartner Hype Cycle for Emerging Technologies*. Retrieved 04/29/2020, from https://www.gartner.com/smarterwithgartner/top-trends-in-the-gartner-hype-cycle-for-emerging-technologies-2017/.

Partel, V., Kakarla, S. C., and Ampatzidis, Y. 2019. "Development and Evaluation of a Low-Cost and Smart Technology for Precision Weed Management Utilizing Artificial Intelligence," *Computers and electronics in agriculture* (157), pp. 339-350.

Peng, Y., Dong, M., and Zuo, M. J. 2010. "Current Status of Machine Prognostics in Condition-Based Maintenance: A Review," *The International Journal of Advanced Manufacturing Technology* (50:1-4), pp. 297-313.

Ramasso, E., and Saxena, A. 2014. "Performance Benchmarking and Analysis of Prognostic Methods for Cmapss Datasets," *International Journal of Prognostics Health Management* (5), pp. 1-15.

Raouf, A., Duffuaa, S., Ben-Daya, M., Dhillon, B., and Liu, Y. 2006. "Human Error in Maintenance: A Review," *Journal of Quality in Maintenance Engineering* (12:1), pp. 21-36.

Ras, G., van Gerven, M., and Haselager, P. 2018. "Explanation Methods in Deep Learning: Users, Values, Concerns and Challenges," in *Explainable and Interpretable Models in Computer Vision and Machine Learning*. Cham: Springer, pp. 19-36.

Richter, H. 2012. "Multiple Sliding Modes with Override Logic: Limit Management in Aircraft Engine Controls," *Journal of Guidance, Control, and Dynamics* (35:4), pp. 1132-1142.

Rudin, C. 2019. "Please Stop Explaining Black Box Models for High Stakes Decisions," *Nature Machine Intelligence* (1:5), pp. 206-215.

Saaty, T. L. 2008. "Decision Making with the Analytic Hierarchy Process," *International Journal of Services Sciences* (1:1), pp. 83-98.

Samek, W., Wiegand, T., and Müller, K.-R. 2017. "Explainable Artificial Intelligence: Understanding, Visualizing and Interpreting Deep Learning Models," *arXiv preprint arXiv:1708.08296*.

Saxena, A., Goebel, K., Simon, D., and Eklund, N. 2008. "Damage Propagation Modeling for Aircraft Engine Run-to-Failure Simulation," *International Conference on Prognostics and Health Management*, Denver, pp. 1-9.

Siau, K., and Wang, W. 2018. "Building Trust in Artificial Intelligence, Machine Learning, and Robotics," *Cutter Business Technology Journal* (31:2), pp. 47-53.

Simonovic, S. P. 1999. "Decision Support System for Flood Management in the Red River Basin," *Canadian Water Resources Journal* (24:3), pp. 203-223.

Stefani, K., and Zschech, P. 2018. "Constituent Elements for Prescriptive Analytics Systems," in: *European Conference on Information System (ECIS),* Stockholm, pp. 1-16

Turek, M. 2016. "Explainable Artificial Intelligence (XAI)," *Defense Advanced Research Projects Agency (DARPA)*. Retrieved 04/29/2020, from https://www.darpa.mil/program/explainable-artificial-intelligence.

Veldman, J., Wortmann, H., and Klingenberg, W. 2011. "Typology of Condition Based Maintenance," *Journal of Quality in Maintenance Engineering* (17:2), pp. 183-202.

Venkatesh, V., Morris, M. G., Davis, G. B., and Davis, F. D. 2003. "User Acceptance of Information Technology: Toward a Unified View," *MIS Quarterly* (27:3), pp. 425-478.

Wang, J., Ma, Y., Zhang, L., Gao, R. X., and Wu, D. 2018. "Deep Learning for Smart Manufacturing: Methods and Applications," *Journal of Manufacturing Systems* (48), pp. 144-156.

Wang, Z., Tang, W., and Pi, D. 2017. "Trajectory Similarity-Based Prediction with Information Fusion for Remaining Useful Life," *International Conference on Intelligent Data Engineering and Automated Learning*, Cham, pp. 270-278.

Wanner, J., Herm, L.-V., and Janiesch, C. 2020. "How Much Is the Black Box? The Value of Explainability in Machine Learning Models," *European Conference on Information Systems (ECIS)*, pp. 1-14.

Yang, F., Du, M., and Hu, X. 2019. "Evaluating Explanation without Ground Truth in Interpretable Machine Learning," *arXiv preprint arXiv:1907.06831*.

Yang, H.-d., and Yoo, Y. 2004. "It's All About Attitude: Revisiting the Technology Acceptance Model," *Decision Support Systems* (38:1), pp. 19-31.

Zheng, S., Ristovski, K., Farahat, A., and Gupta, C. 2017. "Long Short-Term Memory Network for Remaining Useful Life Estimation," *International Conference on Prognostics and Health Management (ICPHM)*, Allen, pp. 88-95.

Zhu, J., Liapis, A., Risi, S., Bidarra, R., and Youngblood, G. M. 2018. "Explainable Ai for Designers: A Human-Centered Perspective on Mixed-Initiative Co-Creation," *Conference on Computational Intelligence and Games (CIG)*, Maastricht, pp. 1-8.

Zschech, P. 2018. "A Taxonomy of Recurring Data Analysis Problems in Maintenance Analytics," *European Conference on Information Systems (ECIS)*, Portsmouth, pp. 1-16.

Zschech, P., Bernien, J., and Heinrich, K. 2019. "Towards a Taxonomic Benchmarking Framework for Predictive Maintenance: The Case of Nasa's Turbofan Degradation," in: *International Conference on Information Systems (ICIS)*. Munich, pp. 1-15.