Electronic Markets
https://doi.org/10.1007/s12525-022-00603-6

RESEARCH PAPER

Decision support for efficient XAI services - A morphological analysis, business model archetypes, and a decision tree

Jana Gerlach[1] · Paul Hoppe[1] · Sarah Jagels[1] · Luisa Licker[1] · Michael H. Breitner[1]

Received: 31 May 2022 / Accepted: 12 October 2022

Abstract

The black-box nature of Artificial Intelligence (AI) models and their associated explainability limitations create a major adoption barrier. Explainable Artificial Intelligence (XAI) aims to make AI models more transparent to address this challenge. Researchers and practitioners apply XAI services to explore relationships in data, improve AI methods, justify AI decisions, and control AI technologies with the goals to improve knowledge about AI and address user needs. The market volume of XAI services has grown significantly. As a result, trustworthiness, reliability, transferability, fairness, and accessibility are required capabilities of XAI for a range of relevant stakeholders, including managers, regulators, users of XAI models, developers, and consumers. We contribute to theory and practice by deducing XAI archetypes and developing a user-centric decision support framework to identify the XAI services most suitable for the requirements of relevant stakeholders. Our decision tree is founded on a literature-based morphological box and a classification of real-world XAI services. Finally, we discussed archetypical business models of XAI services and exemplary use cases.

Keywords Artificial intelligence · Explainability · Morphological analysis · Business models · Archetypes · Decision tree

JEL classification M150 · M210

Motivation and research needs

Artificial Intelligence (AI) has potentially far-reaching applications that can influence people's private and professional lives (Meske et al., 2022). These include the identification of diseases (Aignostics[1]; Meske et al., 2022), job recruitment (iVCV[2]; Sipior et al., 2021), public security (Intelligent Artifact[3]), and risk assessment when granting loans (Wang et al., 2019; ZEST AI[4]). The models used in these instances are often highly complex black boxes (Adadi & Berrada, 2018), meaning that the ability to understand the models' underlying AI processes—and thus the reasons for their decisions—is severely limited. This is problematic because the comprehensibility, explainability, and justification of decisions are of great importance for many applications, including in the health, finance, and energy sectors (Meske et al., 2022).

Although AI is already used for a wide range of activities and provides various benefits, many decision-makers such as managers and executive board members, remain reluctant to integrate AI technologies caused by a limited understanding (Barredo Arrieta et al., 2020). This issue can be addressed by Explainable Artificial Intelligence (XAI) methods, which emphasize the need to make complex models and algorithms understandable and reproducible to humans (Meske et al., 2022). According to Gilpin et al. (2018), the term "explainability" refers to "models that are able to summarize the reasons for neural network behavior, gain the trust of users, or produce insights about the causes of their decisions" (p.80). Moreover, it is not sufficient to gain the trust and

Responsible Editor: Mathias Klier

✉ Jana Gerlach
gerlach@iwi.uni-hannover.de

[1] Institut für Wirtschaftsinformatik, Leibniz Universität Hannover, Königsworther Platz 1, 30167 Hannover, Germany

[1] http://www.aignostics.com.
[2] https://ivcv.eu/.
[3] https://www.intelligent-artifacts.com/
[4] https://zest.ai/.

understanding of model users; the trust and understanding of other relevant stakeholders such as managers, regulators, AI developers, and users or people affected by model decisions are also necessary (Barredo Arrieta et al., 2020).

XAI research includes feature engineering (Wambsganss et al., 2021), algorithmic development and testing (Förster et al., 2021; Xie et al., 2022; Zschech et al., 2022), risks and opportunities (Meske et al., 2022), principles for ethical AI utilization (HLEG-AI, 2019; Seppälä et al., 2021; Thiebes et al., 2021), and the adoption, trust, and usage behavior of XAI (Hamm et al., 2021; Hemmer et al., 2022; Lockey et al., 2021; Stroppiana Tabankov & Möhlmann, 2021). XAI services, both from startups and from established companies like Google, increasingly appear in electronic marketplaces. They offer value creation through applications such as image annotation for healthcare or botanical purposes (Zegami[5]), fraud detection for cybersecurity (Fiddler[6]), or decision support for financial investments (Google Cloud[7]; ZEST AI[8]). In such applications, dashboards, what-if scenarios, and no-code models propose to provide explainability and justified decision-making (DataRobot[9]; Google Cloud[7]; Lagoon[10]).

The worldwide revenue of the XAI market was valued at 4.4 billion U.S. dollars in 2021, and it is forecasted to grow by 2030 to a volume of 21 billion U.S. dollars (Statista, 2022). These XAI services vary widely in terms of target group, purpose, utilized model, and degree of explainability. While several approaches to classify the various design options can be found in literature (e.g., Adadi & Berrada, 2018; Barredo Arrieta et al., 2020), there is a lack of literature-based overviews of existing design options for XAI models in connection with real-world, commercially available XAI services. Thus, we address the following research question (RQ):

RQ1: *Which XAI design options can be extracted from the literature using a morphological analysis?*

To address RQ1, we perform a morphological analysis following Ritchey (2011) and Zwicky (1967) to conceptualize the existing approaches in the literature. We create a morphological box (MBox) as the result of the morphological analysis and structure all the design options of XAI services according to their dimensions and characteristics (Ritchey, 2011).

Haag et al. (2022) observed that many companies are unable to exploit the full potential of AI methods for corporate processes due to the lack of knowledge about AI methods, their application areas, and their possible benefits. Therefore, stakeholders need decision support to identify suitable design options, use cases, and business models; such support, especially when employing XAI solutions and services, reduces AI's entrance threshold. Real-world XAI services are provided by companies specialized in data science. These companies offer commercially available complete XAI solutions or XAI cloud platforms, such as Dataiku,[11] DataRobot[9], ZEST AI[8], and this leads us to our second RQ:

RQ2: *Which archetypical business models can be deduced from classifying real-world XAI services, and how can XAI stakeholders be supported in selecting suitable XAI services for their requirements?*

To address RQ2, we apply the conceptual MBox and classify 40 real-world XAI services into the dimensions and characteristics of our MBox and deduce archetypical XAI business models. This allows us to compare literature and practice and to develop a decision support framework. The latter takes the form of a decision tree for decision-makers and other relevant stakeholders in companies and organizations, thus allowing them to integrate XAI solutions and services into their corporate processes.

RQ1 addresses the current understanding of XAI design options in the literature, and RQ2 focuses on the description and distribution of XAI design options in industrial applications, here in the context of XAI real-world services. A comparison of the literature-based morphological analysis and the archetype analysis of real-world XAI services could reveal possible similarities and differences concerning XAI design options in theory and practice. We thus address a third RQ:

RQ3: *What are the differences between XAI design options in theory and practice, and what impact might these differences have?*

This paper is organized as follows. First, we describe the theoretical background and our research design. Based on this, we present the morphological analysis and assess to what extent it matches real-world XAI applications. Then, we perform a cluster analysis and deduce seven archetypical business models. Finally, we develop a decision support framework as a decision tree; discuss our results, findings, and limitations; and derive recommendations for further research and practice.

[5] https://zegami.com/.

[6] https://fiddler.ai/.

[7] https://cloud.google.com/explainable-ai?hl=de.

[8] https://zest.ai/.

[9] https://www.datarobot.com/.

[10] https://www.data-lagoon.com/.

[11] https://www.dataiku.com/de/.

Theoretical background

According to Kaplan and Haenlein (2019), AI is a "system's ability to correctly interpret external data, learn from such data, and use those learnings to achieve specific goals and tasks through flexible adaptation" (p. 15). The use of AI is expected to grow rapidly in the coming years; according to forecasts, the market for AI software will reach a global revenue of $126 billion by 2025 (Omdia, 2021). AI influences both private lives and entire business models affecting decision-making in areas such as the healthcare, finance, and energy sectors (Haenlein & Kaplan, 2019). Tasks performed by AI can either complement or replace human work (Meske et al., 2022). AI often generates predictions and provides recommendations based on large amounts of data (Kibria et al., 2018). However, the diversity of AI's potential applications raises the question of which decisions should be made by AI models and which should not (Haenlein & Kaplan, 2019). The question is difficult because AI can result in a range of benefits, challenges, and risks, all of which must be weighed against each other (Meske et al., 2022). Due to computers' rapidly increasing processing capacities, high-performance AI systems are possible, today. Indeed, in some cases such as breast cancer detection, the performance of AI exceeds that of humans, underscoring its utility (McKinney et al., 2020). This is especially true for Machine Learning (ML) approaches that use Artificial Neural Networks (ANN). However, AI models of Deep Learning like ANNs are complex and increasingly opaque. Such models are referred to as "black-box models" (Barredo Arrieta et al., 2020). The term refers to models for which it is difficult to understand the internal training and operation of the algorithm, making it a challenge to interpret how the algorithm's outputs are obtained (Adadi & Berrada, 2018).

A potential risk of AI usage is the bias, it can introduce in various forms, including automation, discrimination, and statistical bias (Meske et al., 2022). Automation bias refers to the tendency to over-rely on decisions made by a computer system even when personal decisions are more correct (Goddard et al., 2012; Meske et al., 2022). For example, medical doctors may make decisions based on AI results even though they would have made a different diagnosis without AI. This type of bias arises because humans often tend to accept the recommendations of decision support systems without critically questioning them (Goddard et al., 2012). Meanwhile, discrimination bias can involve, for example, racial or gender discrimination because human bias is present in training data (e.g., text and web corpus; Caliskan et al., 2017; Meske et al., 2022). Finally, statistical bias is the potential distortion between results calculated using historical data and actual data (Meske et al., 2022).

XAI addresses these challenges and risks (Adadi & Berrada, 2018). The goal of addressing the lack of trust and transparency associated with AI models is a major contributor to the emergence of XAI (Adadi & Berrada, 2018; Gilpin et al., 2018; Lipton, 2018). According to Barredo Arrieta et al. (2020), the drivers of XAI depend on the individual stakeholders. For example, users of a model, such as physicians and insurance agents, want to trust an AI model and gain scientific knowledge; regulators want to certify an AI model's compliance with applicable legislation; managers want to evaluate regulatory compliance and understand enterprise AI applications; data scientists, developers, and product owners want to improve product efficiency and develop new functionalities; and people affected by AI model decisions want to understand their situation and verify the fairness of decisions (Barredo Arrieta et al., 2020).

To create explainability, the two goals of interpretability and completeness are addressed. However, it is challenging to simultaneously achieve both goals (Gilpin et al., 2018). Completeness refers to the system's accuracy (i.e., the accuracy of the model; Gilpin et al., 2018), while interpretability describes whether the reasons behind the decision are directly understandable to humans without further explanation (Doshi-Velez & Kim, 2017; Gilpin et al., 2018; Guidotti et al., 2019). However, according to Lipton (2018), "interpretability does not reference a monolithic concept" (p. 42). This means that different AI approaches (e.g., linear regression, Bayesian models, support vector machines, or multi-layer neuronal networks) achieve different levels of explainability through either transparent AI models, which are explainable by design models, or post-hoc models, that provide explainable information on the already developed model (Barredo Arrieta et al., 2020). While post-hoc models often fail to provide insights into exactly how a model works, they may nonetheless provide valuable information for practitioners and users of ML (Lipton, 2018). According to Gilpin et al. (2018), "given the purpose and type of explanation, it is not obvious what the best type of explanation metric is and should be" (p. 88).

There are many ways to achieve XAI, each of which provides different levels and understandings of interpretability (Doshi-Velez & Kim, 2017; Gilpin et al., 2018; Kim, 2018; Lipton, 2018). Adadi and Berrada (2018), Gilpin et al. (2018), and Guidotti et al. (2019) illustrated the variety of XAI techniques, including model distillation, layer-wise relevance propagation, surrogate models, and feature importance. They also discussed these techniques associated with the global or local scope of interpretability and their post-hoc or by-design explanations. Our research shows XAI models' design options and their explainability targets, and we also demonstrate the relative prevalence of these design options in the real world.

Research design and research methods

The research questions address a complex problem concerning decision support for interested stakeholders in XAI design, development, and application. Our research design is structured into three phases: morphological analysis, classification and clustering, and decision support framework development. We present our research procedure in Table 1, then describe it step by step in the following section.

Phase 1

In the first phase, we performed a morphological analysis to identify design options for XAI applications. This analysis allowed us to structure and conceptualize our research topic within the literature, reducing the complexity of the multi-dimensional problem and identifying the interplay between dimensions and characteristics (Ritchey, 2011). This builds the theoretical foundation for the archetype identification and the development of the decision support framework, but it can also contribute to literature and practice on its own.

The first step of morphological analysis is to search and review the relevant literature to identify the dimensions and their specific characteristics. In the context of MBox development, it is mandatory that only one characteristic in each dimension be selected (Ritchey, 2011).

We conducted a systematic literature review in line with Templier and Paré (2015), Watson and Webster (2020), Webster and Watson (2002), and vom Brocke et al. (2015). We browsed the academic databases IEEEXplore, AISeLibary, Science Direct, Springer Link, and ACM Digital Library and searched for articles containing the following keywords in the title or abstract: "XAI" OR "explainable AI" OR "explainable artificial intelligence" AND "taxonomy" OR "framework" OR "components" OR "design" OR "design options" AND "business" OR "business model" OR "service" AND "methods" OR "system" OR "model." Articles had to be peer-reviewed and published between 2017 and 2022.

The keyword-based database search identified 203 scientific publications. After our screening, we excluded all publications not focusing on XAI design options or frameworks. Twelve papers remained after these exclusions. Furthermore, we performed a backward, forward, author, and Google Scholar similarity search with the most important articles in the keyword-based literature search (the most important are, e.g., Adadi & Berrada, 2018; and Barredo Arrieta et al., 2020). After full-text screening, we reached saturation in the search process of scientific publications because no significant novel XAI design options were found. We stopped when we identified ten additional publications. Thus, we included 22 scientific publications for the second step of creating our MBox. All 22 publications can be found in the MBox as references.

Phase 2

To evaluate the theoretical and literature-based MBox, we classified 40 real-world XAI services within the MBox's dimensions and characteristics. This also allowed us to create a data set for our archetype analysis and for our development of the decision support framework. Our data set consists of the 40 XAI services on the vertical axis, while the horizontal axis defines the characteristics and their corresponding dimensions from the MBox. Each XAI service is then checked to determine which characteristics match each dimension. Only one characteristic can be selected for each dimension, see online Appendix A (https://osf.io/b8r7j/?view_only=2a5e19822eb34b618a1ee219936576a7).

To find the real-world XAI services, we used the database crunchbase.com, which provides business and corporate information related to technology companies (Weking et al., 2020), and the search engine Google. We searched for the following keywords: "XAI" OR "explainable AI" OR "explainable artificial intelligence" AND "services" OR "applications" OR "solutions" OR "companies" OR "start-ups." This search identified 78 companies offering XAI services in various disciplines. Due to insufficient information on the companies' websites, we excluded 38 companies. Finally, we classified 40 XAI companies according to the dimensions and characteristics of the MBox and constructed a vector for each examined object along the dimensions.

In Step 4, we conducted a cluster analysis, which allowed us to discover a structure, identify patterns in the data set, and group the classified real-world XAI services, see online Appendix A (https://osf.io/b8r7j/?view_only=2a5e19822eb34b618a1ee219936576a7). According to Kundisch et al. (2021), cluster analysis can be conducted as an evaluation of the MBox with the goal of "better describing, identifying, classifying, analyzing, and clustering objects that represent a certain phenomenon compared to doing so without a taxonomy or other classification schemes" (p. 9). Similar XAI services with similar classified characteristics according to our MBox are grouped into one cluster (Kaufman & Rousseeuw, 1990). We applied the k-means algorithm to cluster the data set with a predefined number of clusters (Kaufman & Rousseeuw, 1990). The k-means algorithm (see online Appendix B (https://osf.io/b8r7j/?view_only=2a5e19822eb34b618a1ee219936576a7)) is an established partitional-based clustering method which has the advantage that it clusters the data set based on its centroid and distance in a simple way. However, the number of centroids, which means the number of clusters, must be a priori-determined (Saputra et al., 2020). To find the optimal number of clusters we used the elbow and silhouette methods (Punj & Stewart, 1983; Rousseeuw, 1987).

Table 1 Research design and research methods

	Phase 1. Morphological analysis		Phase 2. Classification and clustering		Phase 3. Decision support framework development
Steps	Step1: Literature analysis	Step 2: Development of the MBox	Step 3: Classification of real-world XAI services and evaluation	Step 4: Cluster analysis and archetype deduction	Step 5: Development of the decision tree
Tasks	1.1 Keyword-based literature search 1.2 Backward, forward, author, and Google Scholar similarity search 1.3 Systematic literature analysis	2.1 Identification of meta-dimensions 2.2 Identification of dimensions and characteristics	3.1 Data set creation 3.2 Classification of real-world XAI services according to the MBox	4.1 Pre-process data set for clustering 4.2 Run algorithm 4.3 Find an optimal number of clusters 4.4 Derive archetypes	5.1 Split data set into training data and test data 5.2 Run algorithm and select decision tree
Method and reference	Literature review (Templier & Paré, 2015; vom Brocke et al., 2015; Watson & Webster, 2020; Webster & Watson, 2002)	Morphological analysis (Ritchey, 2011; Zwicky, 1967)	Advanced search with Crunchbase (crunchbase.com) and Google search engine	Cluster analysis (Kaufman & Rousseeuw, 1990) via RStudio. Evaluation (Kundisch et al., 2021)	Decision tree development (Pedregosa et al., 2011) via sci-kit learn
Data	*IEEE*Xplore, AISeLibary, Science Direct, SpringerLink, and ACM Digital Library	XAI literature	MBox and real-world XAI services	Data set of classified XAI services	Data set of classified XAI services and results of the cluster analysis
Results and findings	State-of-the-art literature base	Design options of XAI solutions	Data set	Archetypes of XAI business models	Decision support framework for XAI stakeholders

These methods allowed us to determine how close the data is with others within a cluster and how far away one cluster is from the others (Saputra et al., 2020). Based on the clustering results, we derived our archetypical patterns of XAI business models, which involves identifying the similarities among the focuses of our archetypes.

Phase 3

In the third phase, we developed the decision support framework in the form of a decision tree. This framework provides decision support in selecting the most suitable XAI business model for relevant stakeholders. A decision tree is a helpful tool for decision-making in relation to the previously identified XAI business models and archetypes. It is easy to understand and easy to use. Due to the clear structure of the tree with its root nodes, the tree offers decision rules and clearly indicates dependencies (Kamiński et al., 2018). The decision tree can serve as a support framework for decision-makers such as managers, product owners, and data scientists purchasing or programming novel XAI products. The multitude of options to integrate explainability into AI models can be overwhelming, and it is a major responsibility for many decision-makers. Different XAI model requirements can be queried using our decision tree, and based on these answers, a recommendation will be made regarding which XAI business model and archetype should be selected.

We implemented the decision tree algorithm using the Python-based ML toolbox from sci-kit learn (Pedregosa et al., 2011). The archetypes are the recommendations of the decision tree, while the selected characteristics are the respective data features that the model obtains for training. This creates individual vectors of the 40 XAI services (see online Appendix C (https://osf.io/b8r7j/?view_only=2a5e19822eb34b618a1ee219936576a7)). The archetypes are the output that the decision tree tries to predict. The decision tree algorithm (see online Appendix G (https://osf.io/b8r7j/?view_only=2a5e19822eb34b618a1ee219936576a7)) produces binary questions, such as "Will the explainability be integrated into the AI model post-hoc or not?" We manually transformed the answer "or not" to other possible answers extracted from the MBox and the cluster analysis results. To make the decision tree more useful, we added archetype-specific design recommendations that XAI developers can follow to select the best-suited services for their requirements.

Literature review and morphological analysis

Based on the systematic literature review, we identified four publications that classify XAI models in general: Adadi and Berrada (2018), Barredo Arrieta et al. (2020), Li et al., (2020), and Mohseni et al. (2021). These papers served as a basic framework for the development of the MBox and were supplemented by additional topic-specific papers. This allowed us to classify the MBox into three layers: objectives, classification of XAI methods, and XAI methods, see Table 2. The objective layer addresses the target of integrating XAI. This includes the motivation for XAI (e.g., Adadi & Berrada, 2018; Meske et al., 2022) and the goals of several types of XAI users, including AI novices, data experts, and AI experts (Mohseni et al., 2021). Both the classification of XAI methods layer and the XAI methods layer originate from Adadi and Berrada (2018), as they group XAI strategies into classification of the XAI methods according to the complexity, scope, and dependency level of the AI model. XAI methods concern concrete XAI techniques, such as visualization and example-based explanations (Adadi & Berrada, 2018).

Objectives

Explainability can be incorporated into advanced AI solutions for various reasons (D_1). Adadi and Berrada (2018) classified these reasons as follows: *explain to justify* ($C_{1,1}$), which entails a fair decision-making process; *explain to control* ($C_{1,2}$), which develops a better understanding of the algorithm; *explain to improve* ($C_{1,3}$) the algorithm; and *explain to discover* ($C_{1,4}$), which examines the relationships between data.

Once a company has decided to develop an XAI application for one or more of the motivations (*multiple* $C_{1,5}$), various goals for different user groups can be pursued during its implementation: AI novice goals (D_2), data expert goals (D_3), and AI expert goals (D_4). AI novices refer to end-users who apply AI technologies in their everyday lives but have a limited understanding of their underlying systems (Mohseni et al., 2021). For AI novices, the goals of *algorithmic transparency* ($C_{2,1}$), *trust and reliance* ($C_{2,2}$), *bias mitigation* ($C_{2,3}$), and *privacy awareness* ($C_{2,4}$) may be relevant to XAI development (Carvalho et al., 2019; Gerlings et al., 2021; Mohseni et al., 2021). Data experts are data scientists or domain experts who use AI to gain insights from data. Though they have a particularly good understanding of their application area, they are unfamiliar with the technical processes required to make AI work. Data experts may be particularly interested in *visualizing and inspecting the models* ($C_{3,1}$) and *tuning and selecting* ($C_{3,2}$) models for specific problems. AI experts, by contrast, are responsible for developing, implementing, and continuously improving AI algorithms and explainability techniques. *Model interpretability* ($C_{4,1}$) is an important criterion for AI experts because it helps them understand the AI's processes for learning from data in general and making decisions in specific contexts. In addition, they use explainability techniques to improve the model and the underlying training process. The *model debugging* ($C_{4,2}$) characteristic captures these goals. In conclusion, companies will pursue different goals depending

Table 2 Morphological box

	Dimension Di	Characteristics $C_{i,j}$		Reference
Objectives	D_1 Motivation for explanation	$C_{1,1}$ Explain to justify $C_{1,3}$ Explain to improve $C_{1,5}$ Multiple	$C_{1,2}$ Explain to control $C_{1,4}$ Explain to discover	Adadi and Berrada (2018), Meske et al. (2022), Thiebes et al. (2021)
	D_2 AI novice goals	$C_{2,1}$ Algorithmic transparency $C_{2,3}$ Bias mitigation $C_{2,5}$ Multiple	$C_{2,2}$ Trust and reliance $C_{2,4}$ Privacy awareness $C_{2,6}$ None	Carvalho et al., (2019) Gerlings et al. (2021), Mohseni et al. (2021)
	D_3 Data expert goals	$C_{3,1}$ Model visualization and inspection $C_{3,3}$ Both	$C_{3,2}$ Model tuning and selection $C_{3,4}$ None	Mohseni et al. (2021)
	D_4 AI expert goals	$C_{4,1}$ Model interpretability $C_{4,3}$ Both	$C_{4,2}$ Model debugging $C_{4,4}$ None	Mohseni et al. (2021)
Classification of XAI methods	D_5 Complexity-related methods	$C_{5,1}$ Post-hoc	$C_{5,2}$ By design	Adadi and Berrada (2018), Alamri and Alharbi (2021)
	D_6 Model-related methods	$C_{6,1}$ Model-specific	$C_{6,2}$ Model-agnostic	Adadi and Berrada (2018), Markus et al. (2021), Rai (2020)
	D_7 Scope-related methods	$C_{7,1}$ Global $C_{7,3}$ Both	$C_{7,2}$ Local	Adadi and Berrada (2018), Guidotti et al. (2019), Ivaturi et al. (2021), Setzu et al. (2021), Rai (2020)
	D_8 Input data types	$C_{8,1}$ Tabular data $C_{8,3}$ Text $C_{8,5}$ Multiple	$C_{8,2}$ Image $C_{8,4}$ Graph data	Li et al., (2020), Linardatos et al. (2021)
XAI methods	D_9 Explanation by influence method	$C_{9,1}$ Sensitivity analysis $C_{9,3}$ Feature importance $C_{9,5}$ None	$C_{9,2}$ LRP $C_{9,4}$ Multiple	Barredo Arrieta et al. (2020), Adadi and Berrada (2018), Li et al., (2020), Markus et al. (2021), Meister et al. (2021), Vilone and Longo (2021), Zhang et al. (2021)
	D_{10} Visual explanation	$C_{10,1}$ PDP $C_{10,3}$ Feature relevance visualization $C_{10,5}$ None	$C_{10,2}$ ICE $C_{10,4}$ Multiple	Adadi and Berrada (2018), Barredo Arrieta et al. (2020), Curia (2021), Li et al., (2020)
	D_{11} Explanation by simplification	$C_{11,1}$ Rule extraction $C_{11,3}$ Surrogate model	$C_{11,2}$ Model distillation $C_{11,4}$ None	Adadi and Berrada (2018), Barredo Arrieta et al. (2020), Li et al., (2020), Kridel et al. (2020), Wastensteiner et al. (2021)
	D_{12} Example-based explanations	$C_{12,1}$ Prototypes and criticisms $C_{12,3}$ Both	$C_{12,2}$ Counterfactual explanations $C_{12,4}$ None	Adadi and Berrada (2018), Barredo Arrieta et al. (2020), Markus et al. (2021), Stepin et al. (2021)
	D_{13} Text explanations	$C_{13,1}$ Yes	$C_{13,2}$ No	Barredo Arrieta et al. (2020)

on the target group for an XAI application (Mohseni et al., 2021).

Classification of XAI methods

XAI methods can be classified into four dimensions: complexity-related methods (D_5; Adadi & Berrada, 2018), model-related methods (D_6; Adadi & Berrada, 2018; Markus et al., 2021; Rai 2019), scope-related methods (D_7;Adadi & Berrada, 2018 ; Guidotti et al., 2019 ; Ivaturi et al., 2021 ; Setzu et al., 2021), and input data types (D_8; Li et al., 2020; Linardatos et al., 2021).

The first dimension, complexity-related methods, can be divided into *post-hoc* ($C_{5,1}$) and *by design* ($C_{5,2}$) explanations (Adadi & Berrada, 2018; Alamri & Alharbi, 2021; Mohseni et al., 2021). The former occurs in addition to the black-box model, while by design explanations occur during the model's training phase (Alamri & Alharbi, 2021). Following Adadi and Berrada (2018), "the complexity of a machine-learning model is directly related to its interpretability" (p. 52147). According to them, more

complex methods provide less interpretability, and simpler methods are more interpretable. However, there is an ongoing debate concerning the relationship between model complexity and accuracy in literature. To analyze this relationship, Koziol and Weitz (2021) examined various pricing models and input data types (e.g., historical data, solvency data, and product data). They found that under normal circumstances, in their case, a normal market environment, increased model complexity does not necessarily improve its output accuracy and that input data can also play a central role (Koziol & Weitz, 2021). In an earlier evaluation of the complexity and the accuracy of different forecasting models, Ahlburg (1995) concluded that "it is too early to say whether simple models are more accurate than complex models or whether causal models are more accurate than noncausal models" (p. 287); this debate continues today.

The model-related methods can be distinguished into *model-specific* ($C_{6,1}$) and *model-agnostic* ($C_{6,2}$) interpretability techniques. *Model-specific* techniques can only be applied to a certain class of models or algorithms, while *model-agnostic* techniques can be used for any algorithm type (Adadi & Berrada, 2018; Markus et al., 2021; Rai 2019). Moreover, *model-specific* techniques consider only certain model types when specific types of explanation are required. The disadvantage of these techniques is that selecting a model that provides a certain type of explanation often reduces the model's representativeness (Adadi & Berrada, 2018). According to Adadi and Berrada (2018), "*model-agnostic* interpretability techniques are convenient, they often rely on surrogate models or other approximations that can degrade the accuracy of the explanations they provide" (p. 52151). This is not the case for *model-specific* interpretations since they refer to a specific model (Adadi & Berrada, 2018).

There are two variations of the scope of interpretability: *global* ($C_{7,1}$) and *local* ($C_{7,2}$; Adadi & Berrada, 2018). *Local* interpretability means that only one specific decision can be explained. In contrast, *global* interpretability refers to understanding the entire system and the connection between input and output variables so that every decision is comprehensible (Adadi & Berrada, 2018; Guidotti et al., 2019; Ivaturi et al., 2021; Setzu et al., 2021). Though *global* interpretability is useful, it is difficult to implement. Conversely, *local* interpretability is easier to achieve and is commonly used (Adadi & Berrada, 2018).

Determining which method of explainability should be used also depends on the available input data type. Some models can be applied to data in *tabular* ($C_{8,1}$), *image* ($C_{8,2}$), *text* ($C_{8,3}$), or *graphical* ($C_{8,4}$) form (Li et al., 2020; Linardatos et al., 2021). To include the option of choosing more than one data type from ($C_{8,1}$ to $C_{8,4}$), the characteristic *multiple* ($C_{8,5}$) can be selected either.

XAI methods

XAI methods can be classified into five dimensions: explanation by influence (D_9), visual explanations (D_{10}), explanation by simplification (D_{11}), example-based explanations (D_{12}), and text explanations (D_{13}).

The first dimension is explanation by influence methods, which are applied to analyze the relevance or importance of a certain model feature to prediction performance (Adadi & Berrada, 2018; Barredo Arrieta et al., 2020). In the MBox, we identified three characteristics within this dimension: *sensitivity analysis* ($C_{9,1}$), *layer-wise relevance propagation* (*LRP*; $C_{9,2}$), and *feature importance* ($C_{9,3}$). *Sensitivity analysis* aims to determine the influence of input or weight perturbations on the output (Ruck et al., 1990); measures the usefulness of input features, and identifies which feature has the most significant impact on the prediction (Kridel et al., 2020). The second characteristic is *LRP* (Bach et al., 2015), which includes different layers, such as the input, hidden, and output layers of ANNs. Starting at the output layer, a relevance value is calculated for every neuron in each layer depending on the weights, the activation, and the relevance value of the neuron of a deeper layer. In this way, a relevance value can be determined backward up to each network neuron's input layer (Bach et al., 2015); thus, the *LRP* "identifies pivotal properties for the prediction" (Adadi & Berrada, 2018, p. 52150). Finally, *feature importance* methods can provide either local or global explanations. One approach for global explanation is random trees (Breiman, 2001). Local explanations can be provided by Shapley additive explanation (SHAP), which measures each feature's contribution to the prediction (Lundberg & Lee, 2017).

Visual explanation aims to illustrate an AI model's behavior by analyzing the interactions of input features; it is often applied with other techniques to improve users' understanding of the model (Barredo Arrieta et al., 2020). The literature distinguishes between *partial dependence plot* (*PDP*; $C_{10,1}$), *individual conditional expectation* (*ICE*; $C_{10,2}$), and *feature relevance visualization* ($C_{10,3}$; Adadi & Berrada, 2018). *PDPs* visualize the average partial relationship between input variables and the predicted outcome of *post-hoc* interpretable AI algorithms. They can be classified as a *model-agnostic* XAI method that can achieve either *local* or *global* interpretability. In this context, the influence of one or several features on the prediction can be analyzed (Adadi & Berrada, 2018; Hakkoum et al., 2021). The second type of visual explanation is *ICE*, a *model-agnostic* method that enables *local* interpretability. While *PDPs* use the average effect of the feature on the prediction, *ICE* plots disaggregate the *PDP* and focus on specific instances (Adadi & Berrada, 2018). The selected features are modified (perturbed) in an iterative process while all other features remain unchanged (Li et al., 2020). If there are any

interactions, examining average effects can lead to an erroneous estimation of complexity in heterogenous predicted outcomes (Curia, 2021). The third characteristic is *feature importance*, which aggregates several methods to visualize the relevance of specific features of an AI algorithm (Adadi & Berrada, 2018).

The third dimension of the XAI methods layer is explanation by simplification (D_{11}). This includes all XAI concepts that develop completely new, explainable models based on the trained AI model. The objective is to achieve a less complex model while maintaining the same prediction accuracy (Barredo Arrieta et al., 2020). We identified three such methods in the literature search: *rule extraction* ($C_{11,1}$), *model distillation* ($C_{11,2}$), and *surrogate models* ($C_{11,3}$).

In *rule extraction*, the knowledge the ANN gains through training is made explainable by extracting rules that approximate the ANN's decision-making path using input and output data (Adadi & Berrada, 2018; Li et al., 2020). The second method is *model distillation*, which can be classified as a type of model compression. If it is applied to a deep neuronal network, a deep network called teacher is trained with a large data set. If this model performs accurately, its knowledge can be transferred to a less complex model called the student. The technique aims to find a student who mimics the teacher, leading to a better understanding of the complex model while maintaining prediction accuracy (Adadi & Berrada, 2018). The next characteristic is the *surrogate model,* which is *model-agnostic* with either *local* or *global* interpretability (Hakkoum et al., 2021). In general, surrogate representations are approximations of the actual AI models. These approximated models are much simpler, reducing the complexity of the AI algorithm. To achieve an output that is as accurate as possible, surrogate representations are trained on the predictions of the black-box model using methods such as linear regressions. This improves interpretability but can harm prediction performance (Adadi & Berrada, 2018).

The fourth dimension identified is example-based explanation methods (D_{12}). These methods are useful if the distribution of the training data set is complex and difficult to understand (Li et al., 2020). In this dimension, a distinction between *prototypes and criticisms* ($C_{12,1}$) and *counterfactual explanations* ($C_{12,2}$) is made (Adadi & Berrada, 2018). In this context, "a prototype is a representative data instance from the original data set" and "a criticism is a data instance that is not well represented by the set of prototypes" (Li et al., 2020, p. 8). This method can provide insights into the distribution of the original data set, and criticisms are determined by maximizing the difference in the distribution between the data set and the prototype (Li et al., 2020). Finally, *counterfactual explanations* seek "to find the smallest change of the feature value so that it can change the prediction into the desired outcome" (Li et al., 2020, p. 8).

The last identified dimension within the layer of XAI methods is text explanation (D_{13}). Text explanations provide natural language generated through a learning process that explains an AI model's results. Thus, text explanation cannot be seen as a standalone explanation method. Instead, other techniques provide numbers or visualizations as input to the text explanation model, which outputs natural language explanations (Bennetot et al., 2019).

Classification and cluster analysis

To address RQ2, we performed a cluster analysis to identify the archetypical patterns of XAI business models using our MBox and classification results of real-world XAI services. We classified the XAI services by visiting the website of every XAI service provider (see Appendix 1). These websites describe each XAI service and possible use cases. We only included XAI services that directly state that they offer XAI methods. We examined the 40 XAI services with the dimensions and characteristics of our MBox, ensuring that only one characteristic was selected per dimension. All authors simultaneously and independently classified the 40 XAI services to fulfill the four-eyes principle for validated results. In the case of a disagreement about a characteristic's classification, the authors discussed the classification results.

The resulting data set from classifying the 40 XAI services can be found in the online Appendix A (https://osf.io/b8r7j/?view_only=2a5e19822eb34b618a1ee219936576a7). Based on this analysis, we imported the data set into RStudio and clustered it using the k-means algorithm. The RStudio algorithm can be found in online Appendix B (https://osf.io/b8r7j/?view_only=2a5e19822eb34b618a1ee219936576a7). This algorithm allows us to merge XAI services with the same characteristics into a cluster. Specifically, the data set consists of the 40 XAI services on the vertical axis, the MBox dimensions, and corresponding characteristics on the horizontal axis (see online Appendix A (https://osf.io/b8r7j/?view_only=2a5e19822eb34b618a1ee219936576a7)). For each XAI service, we created a row with zeros and ones. For each dimension and each XAI service, only one characteristic can be marked with a one, which denotes the special characteristic. The applied k-means algorithm grouped XAI services with similar marks for the same characteristics. Patterns of similarities and differences can be identified and incorporated into archetypal business models. However, the optimal number of clusters must be identified before clustering. For this, we followed the silhouette method (Punj & Stewart, 1983; Rousseeuw, 1987), which indicated that seven was the optimal number of clusters, see online Appendix D (https://osf.io/b8r7j/?view_only=2a5e19822eb34b618a1ee219936576a7). The graphical output of the elbow method indicated no clear result, see online Appendix E (https://osf.io/b8r7j/?view_only=2a5e19822eb34b618a1ee219936576a7). In addition, the plotted clusters indicated that the seven clusters are separated from each other and, at the same time, are not too small, in the sense that only one XAI service

was contained in a cluster, see online Appendix F (https://osf.io/b8r7j/?view_only=2a5e19822eb34b618a1ee219936576a7).

In addition, we accepted the tradeoff created by this number of clusters between the level of detail and the number of services in each cluster. Interpreting them is difficult if too few clusters are determined because too many XAI services are merged. Nevertheless, a cluster must consist of more than one service. Table 3 visualizes the clustering analysis results and shows the percentage distribution of features in the seven archetypes and the column between the characteristics. The first cluster shows the percentage distribution in all examined XAI services. Each characteristic is color labeled, with 0% in white and 100% in dark gray. For example, dimension D_5, explainability integration, is in Archetype 1 at 100% for the characteristic $C_{5,1}$. For the cluster analysis, we deleted the characteristics $C_{1,1}$, $C_{2,1}$, and $C_{8,4}$ because they did not appear in the real-world XAI services. All examined XAI services, and their assigned archetypes are listed in the online Appendix C (https://osf.io/b8r7j/?view_only=2a5e19822eb34b618a1ee219936576a7).

Archetype 1—XAI to support decision-making

Archetype 1 consists of the three XAI services ZEST AI,[12] DreamQuark,[13] and Spin Analytics,[14] which offer solutions for the financial industry (see their websites or their crunchbase.com descriptions). The goal for AI novices is to increase trust and reliance on AI models, while the goal for AI experts is to interpret and debug the models. This archetype is characterized by a local scope of related explainability, resulting in only model-specific explainability. Furthermore, explainability is added post-hoc to already solved AI models. The method requires tabular data and produces counterfactual, example-based explanations. This business model is characterized by the use cases of credit and risk decisions (Spin Analytics), lending decision-making (ZEST AI), and asset management (DreamQuark).

Archetype 2—XAI to improve corporate metrics

Archetype 2 includes five XAI services—Cycorp,[15] Minerva Intelligence,[16] Stratyfy,[17] Cognino AI,[18] and Corpy & Co.[19]—in the finance, manufacturing, and healthcare sectors. They target improvable workflows to minimize corporate costs and maximize profits. The XAI services in this archetype are model-agnostic and thus have a global scope of explainability. Layer-wise relevance propagation is the primary method used to explain relevant features; no visual or example-based explanations are provided. Rule extraction methods are used to simplify the model and its results.

Archetype 3—XAI for no-code models

Archetype 3 consists of nine XAI services (e.g., Stride[20] and Akkai Kaeru[21]) that offer solutions for the finance and healthcare sectors. This archetype is characterized by providing no- or low-code models to simplify AI models and increase their usability. Explainability is integrated post-hoc for specific models on a local level, and no explanations by simplification or example-based explanations are provided.

Archetype 4—XAI for transparent and trustworthy AI

Archetype 4 consists of six XAI services that strengthen transparency and trust in AI for corporate decision support. The services are not model-agnostic; they add explainability post-hoc, and only specific models are explained. Multiple input data are used, such as videos and photos (iVCV[22]) and financial data (DydonAI[23]). For use cases such as underwriting (xcoring[24]), recruiting (iVCV), and investment banking (SCALNYX[25]), XAI can support strategic decision-making or recruiting by providing transparent results and causal justifications in a less biased way.

Archetype 5—XAI to leverage data

Archetype 5 consists of seven XAI services (e.g., Zegami,[26] RISHI-XAI,[27] and HMX[28]) to improve the value creation of utilized data such as images, videos, and corporate data. XAI services in this business model are explainable by design in a model-agnostic way. To explain the influencing features of the AI model, this business model uses layer-wise relevance propagation. Corporations' daily core activities (e.g., project management) can be carried out more efficiently

[12] https://zest.ai/.
[13] http://www.dreamquark.com/.
[14] http://spin-analytics.com.
[15] http://www.cyc.com.
[16] https://minervaintelligence.com.
[17] http://www.stratyfy.com.
[18] https://www.cognino.ai/.
[19] https://corpy.co/.
[20] http://www.stride.ai/.
[21] https://www.akaikaeru.com.
[22] https://ivcv.eu/.
[23] http://www.dydon.net.
[24] https://www.xcoring.ai/.
[25] https://www.scalnyx.com/.
[26] https://zegami.com/.
[27] https://www.digite.com/.
[28] https://www.hmx.ai/.

Table 3 Results of the cluster analysis[a]

Dimensions		Clusters	All	1	2	3	4	5	6	7
		Characteristics	$n=40$	$n=3$	$n=5$	$n=9$	$n=6$	$n=7$	$n=8$	$n=2$
Objectives	D_1 Motivation for explanation	$C_{1,2}$ Explain to control	5%	33%	0%	0%	0%	0%	12.5%	0%
		$C_{1,3}$ Explain to improve	30%	0%	20%	78%	50%	14%	0%	0%
		$C_{1,4}$ Explain to discover	12.5%	0%	20%	11%	17%	0%	0%	100%
		$C_{1,5}$ Multiple	52.5%	67%	60%	11%	33%	86%	87.5%	0%
	D_2 AI novice goals	$C_{2,2}$ Trust and reliance	12.5%	67%	20%	0%	0%	0%	25%	0%
		$C_{2,3}$ Bias mitigation	20%	0%	40%	33.3%	33%	0%	0%	50%
		$C_{2,4}$ Privacy awareness	10%	0%	0%	33.3%	17%	0%	0%	0%
		$C_{2,5}$ Multiple	57.5%	33%	40%	33.3%	50%	100%	75%	50%
	D_3 Data expert goals	$C_{3,1}$ Model visualization and inspection	60%	33.3%	80%	89%	50%	29%	50%	100%
		$C_{3,2}$ Model tuning and selection	2.5%	33.3%	0%	0%	0%	0%	0%	0%
		$C_{3,3}$ Both	30%	0%	20%	0%	33%	71%	50%	0%
		$C_{3,4}$ None	7.5%	33.3%	0%	11%	17%	0%	0%	0%
	D_4 AI expert goals	$C_{4,1}$ Model interpretability	67.5%	67%	80%	67%	83%	86%	25%	100%
		$C_{4,2}$ Model debugging	2.5%	33%	0%	0%	0%	0%	0%	0%
		$C_{4,3}$ Both	17.5%	0%	0%	0%	0%	14%	75%	0%
		$C_{4,4}$ None	12.5%	0%	20%	33%	17%	0%	0%	0%
Classification of XAI methods	D_5 Explainability integration	$C_{5,1}$ Post-hoc	52.5%	100%	60%	100%	100%	0%	0%	0%
		$C_{5,2}$ By design	47.5%	0%	40%	0%	0%	100%	100%	100%
	D_6 Model-related methods	$C_{6,1}$ Model-specific	52.5%	100%	0%	100%	100%	0%	12%	100%
		$C_{6,2}$ Model-agnostic	47.5%	0%	100%	0%	0%	100%	88%	0%
	D_7 Scope-related methods	$C_{7,1}$ Global	52.5%	0%	100%	11%	50%	71%	75%	50%
		$C_{7,2}$ Local	37.5%	100%	0%	89%	50%	0%	0%	50%
		$C_{7,3}$ Both	10%	0%	0%	0%	0%	29%	25%	0%
	D_8 Input data types	$C_{8,1}$ Tabular data	17.5%	100%	0%	33%	0%	0%	12.5%	0%
		$C_{8,2}$ Image	5%	0%	0%	11%	0%	0%	0%	50%
		$C_{8,3}$ Text	2.5%	0%	0%	0%	0%	0%	12.5%	0%
		$C_{8,5}$ Multiple	75%	0%	100%	56%	100%	100%	75%	50%

Table 3 (continued)

Dimensions		Clusters Characteristics	All $n=40$	1 $n=3$	2 $n=5$	3 $n=9$	4 $n=6$	5 $n=7$	6 $n=8$	7 $n=2$
XAI methods	D_9 Explanation by influence method	$C_{9,1}$ Sensitivity analysis	25%	33%	20%	67%	33%	0%	0%	0%
		$C_{9,2}$ LRP	35%	0%	60%	0%	50%	100%	0%	50%
		$C_{9,3}$ Feature importance	2.5%	0%	0%	0%	17%	0%	0%	0%
		$C_{9,4}$ Multiple	27.5%	67%	0%	0%	0%	0%	100%	50%
		$C_{9,5}$ None	10%	0%	20%	33%	0%	0%	0%	0%
	D_{10} Visual explanation	$C_{10,1}$ PDP	35%	33.3%	0%	33%	67%	86%	0%	0%
		$C_{10,2}$ ICE	12.5%	33.3%	0%	0%	33%	14%	12.5%	0%
		$C_{10,3}$ Feature relevance visualization	2.5%	0%	0%	0%	0%	0%	0%	50%
		$C_{10,4}$ Multiple	22.5%	33.3%	0%	0%	0%	0%	87.5%	50%
		$C_{10,5}$ None	27.5%	0%	100%	67%	0%	0%	0%	0%
	D_{11} Explanation by simplification	$C_{11,1}$ Rule extraction	40%	33.3%	80%	0%	100%	57%	12.5%	0%
		$C_{11,2}$ Model distillation	10%	0%	0%	0%	0%	29%	25%	0%
		$C_{11,3}$ Surrogate model	12.5%	33.3%	0%	0%	0%	0%	50%	0%
		$C_{11,4}$ None	37.5%	33.3%	20%	100%	0%	14%	12.5%	100%
	D_{12} Example-based explanations	$C_{12,1}$ Prototypes and criticisms	5%	0%	0%	0%	0%	0%	0%	100%
		$C_{12,2}$ Counterfactual explanations	10%	67%	0%	0%	0%	0%	25%	0%
		$C_{12,3}$ Both	17.5%	0%	0%	11%	0%	0%	75%	0%
		$C_{12,4}$ None	67.5%	33%	100%	89%	100%	100%	0%	0%
	D_{13} Text explanations	$C_{13,1}$ Yes	95%	67%	100%	100%	100%	100%	100%	50%
		$C_{13,2}$ No	5%	33%	0%	0%	0%	0%	0%	50%

[a]Due to rounding, there are deviations from 100% for D_i and the single clusters

by extracting insights from a large volume of historical data to accumulate useful knowledge about future results (RISHI-XAI).

Archetype 6—XAI to democratize data science

Archetype 6 includes eight XAI services (e.g., Fiddler,[29] Dataiku,[30] Beyond Limits[31]) that comprehensively democratize XAI models' results. This archetype is collective in its specifications because it utilizes multiple characteristics. For this reason, we reexamined all eight XAI services in this archetype in detail by using their websites and crunchbase.com descriptions and discovered that the archetype's focus is on the input data and its impact on the AI model. Multiple visualization methods such as dashboards, reports, and what-if-scenarios, e.g., Fiddler, help reduce the time for error correction, improve efficiency and accuracy, and encourage trust in AI technologies and adoption. Use cases include detecting damages on solar cells (HACARUS[32]) and credit risk assessment for lenders or stock selection (Beyond Limits).

Archetype 7—XAI to uncover new insights

This is the smallest archetype. It features only two XAI services (Aignostics[33] and clearbox.ai[34]) and aims to discover data's potential for various purposes. Aignostics provides a diagnosis platform to discover biomarkers in biological images to identify evidence of diseases, while clearbox.ai generates synthetic data to improve data sets or anonymize sensitive data. Prototypes and criticisms are utilized in this business model to explain data features by employing examples, but no explanation by simplification is provided.

Discussion and a decision support framework

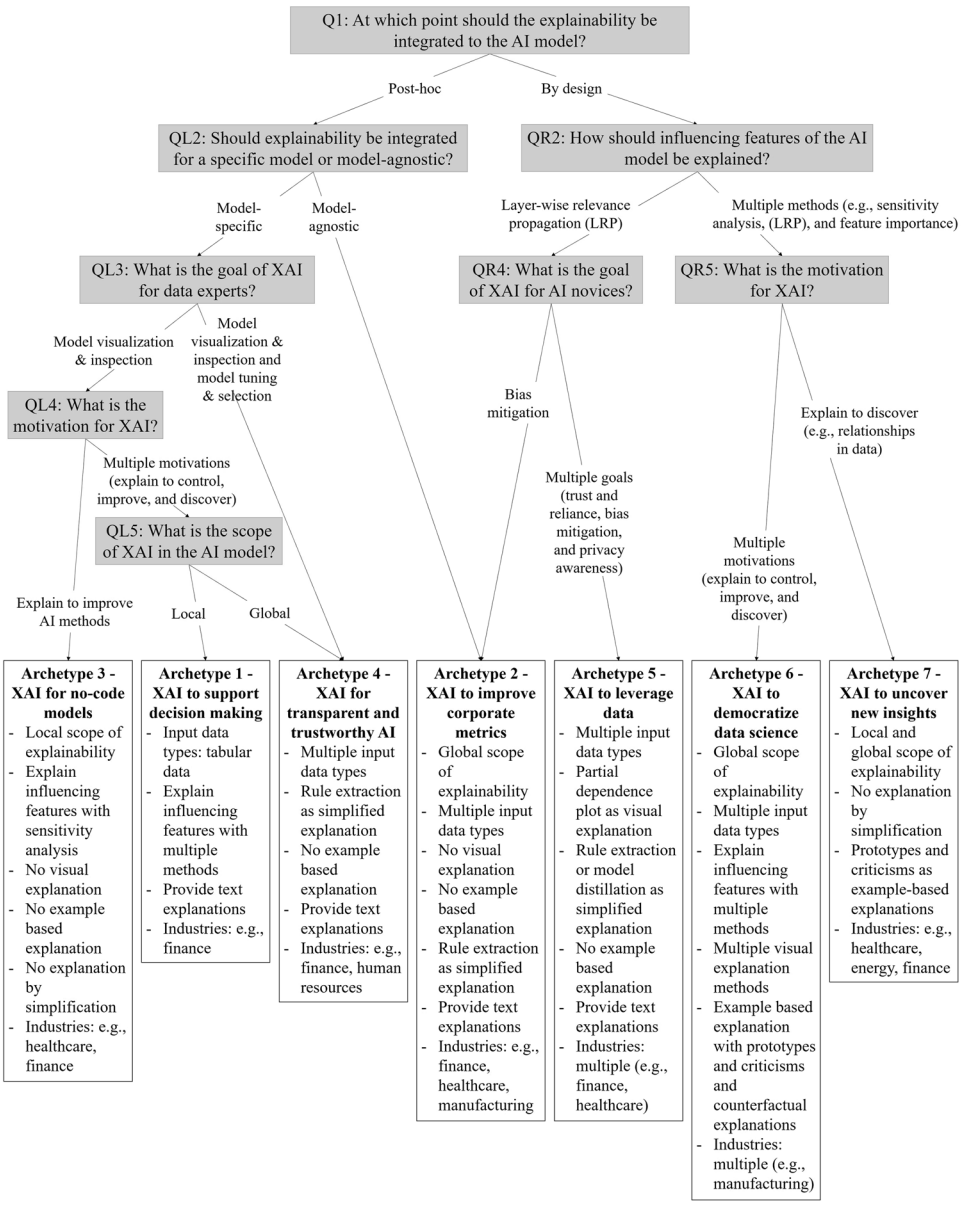
Researchers and practitioners are interested in XAI business models to help them explore data relationships, improve AI methods, justify AI decisions, and control XAI technologies while simultaneously meeting user needs (Adadi & Berrada, 2018; Meske et al., 2022; Thiebes et al., 2021). In contrast, many other scientists have focused on XAI algorithms and proposed artifacts to increase unbiased AI decision-making (Xie et al., 2022) or to understand the behavior of an AI system (Polzer et al., 2022). To benefit from such solutions, users and interested stakeholders such as managers, data scientists, and AI developers must determine which XAI solution best fits their requirements.

According to Haag et al. (2022), though ML has been applied successfully in various contexts, its effectiveness remains constrained by firms' limited knowledge of its possible uses. To address this limitation and RQ2, we developed a decision support framework to help stakeholders select XAI business models and design elements according to their needs for explainability and value creation. Our decision tree provides a market overview and clarifies the XAI selection process by asking binary questions whose answers lead to a specific archetypical business model. Since the archetypes have different purposes and methods, a particular business model can fit a certain decision-makers requirement better than others. In addition, the variety of XAI design options can be challenging and complex, so the decision tree helps to provide an initial overview. As described above in Phase 3 of the research method, we generated the decision tree based on the data set we created by classifying real-world XAI services within the literature-based MBox. Our classified dimensions and characteristics are the decision features, while the archetypes are the respective decision classes produced at the end of the decision tree. The final decision tree is illustrated in Fig. 1. Only questions are included that can be answered at the beginning of a planning phase for an XAI service, and a maximum of five questions are required to determine which business model and archetype to adopt.

To explain the decision tree, the left path will be described. Q1 asks at which point explainability should be integrated into the AI model and can be answered with post-hoc or by design explainability. This question divides the possible business models into two paths. QL2 asks whether XAI services should be integrated in a model-specific or model-agnostic way. If the model-agnostic approach is selected, Archetype 2 is recommended. If the model-specific approach is selected, the next question asks which goal should be pursued by XAI services for data experts. Either model visualization and inspection or model visualization and inspection and model tuning can be answered. If the second answer is given, Archetype 4 is recommended. If the first answer is given, the question regarding the motivation for XAI services follows. This question can be answered either by explaining to improve AI models or with multiple motivations. If the first answer is given, then Archetype 3 is recommended. If the second answer is given, then the question regarding the scope of XAI services in the AI model follows. If the scope is global, then Archetype 4 is

[29] https://fiddler.ai/.
[30] https://www.dataiku.com/de/.
[31] https://www.beyond.ai/.
[32] http://hacarus.com/.
[33] http://www.aignostics.com.
[34] https://www.clearbox.ai/.

Fig. 1 Decision tree



recommended, while Archetype 1 is recommended for local scope.

Furthermore, we have noticed that the offer of XAI services is becoming increasingly diverse and that its market volume is predicted to grow significantly (Statista, 2022). From the MBox, we observe that the characteristics identified in the literature can be well classified using real-world XAI services and we do not add additional dimensions and characteristics. From this classification, we defined seven archetypes that can be deduced as business models, which were named according to their contribution to value creation. Hence, we noticed that the benefit accrues to either XAI application users, companies, or customers. "XAI to support decision-making," "XAI for no-code models," "XAI to leverage data," and "XAI to democratize data science" provide the most benefit for users, and they can facilitate and accelerate AI system workflows. Furthermore, Archetypes 5 and 6 can extract more value from the AI models and data sets through techniques such as visualization. "XAI to improve corporate metrics" and "XAI to leverage data" provide the most benefit for companies using XAI services. New data-driven business opportunities can be accessed by leveraging data. Finally, "XAI for transparent and trustworthy AI" and "XAI to uncover new insights" benefit customers most. Customers or people affected by XAI models' decisions can be sure that companies' decisions are justified, for example, in the healthcare sector, patients can benefit from new diagnostic technologies to identify diseases.

With the help of our decision tree, we derived recommendations for decision-makers such as managers, developers,

and data scientists interested in XAI solutions. The archetypes and design options recommended by the decision tree can help decision-makers identify which design options from the MBox should be considered for the particular explanation requirements. In addition, our study can help increase the acceptance and knowledge of regulatory authorities, users, and people affected by XAI models' decisions.

To address RQ3, we mapped real-world XAI services to the MBox characteristics and identified differences in how often XAI methods are offered in practice (Table 3). The explanation by influence method (D_9) is frequently used, as only 10% of the 40 XAI services do not use any methods listed in D_9. In particular, sensitivity analysis ($C_{9,1}$) and LRP ($C_{9,2}$) are often applied ($C_{9,1}$: 25%, $C_{9,2}$: 35%). The method of rule extraction ($C_{11,1}$) from the dimension of explanation by simplification is also frequently used (40%), as are text explanations ($C_{13,1}$; 95%).

In contrast, example-based explanations are used less frequently, including prototypes and criticisms ($C_{12,1}$: 5%) and counterfactual explanations ($C_{12,2}$: 10%). In total, 67.5% of the 40 XAI services do not use these methods. Moreover, the methods of explanation by simplification are not used by ($C_{11,4}$: 37.5%. These include, in particular, model distillation ($C_{11,2}$: 10%) and the surrogate model ($C_{11,3}$: 12.5%), but rule extraction is used more frequently ($C_{11,1}$: 40%).

The methods used in real-world XAI services often offer visualization and graphical representation. These show the influence of the input data changes on the output and, thus, the AI prediction. Meanwhile, methods used less frequently show interpretations of the models or explain their behavior. According to Barocas et al. (2020) and Crupi et al. (2021), example-based explanations are not sufficient to develop feasible measures that a user can apply. This is consistent with the few use cases of real-world XAI services that apply the methods. Archetype 7 is the only archetype to use the methods of prototypes and criticisms; one use case in this archetype is the uncovering of biomarkers for pathologists. The goal here is to uncover patterns that would be difficult to identify visually. The user does not need any further instructions after the event of a discovery. This suggests that XAI services that provide decision support with instructions for action are offered especially frequently; such methods include the modification of input data (e.g., explanation by influence method) as employed by Adadi and Berrada (2018). Services that explain the output of the AI model are offered less often; such models can only show that something should be changed but do not indicate how. A mixture of both approaches could achieve the best balance between service levels in terms of explanation and decision support. However, if the focus is on decision support, this may raise the risk of losing the explanation of how an AI model works. The acceptance of the regulatory authorities, users, and customers who are affected by AI results can decrease as a result. Therefore, to build and stabilize acceptance, it is important to pursue both explanation and decision support.

By examining real-world XAI services, we were able to determine that the group of private persons or end-consumers was not targeted except by the DataRobot service.[35] In the case of DataRobot, decisions in the consumer domain can be enriched with explanations, but this is only because of the tool's universal applicability and is not explicitly described in a use case. The reasons why DataRobot considers explainability to be particularly relevant to this domain are not provided. This may be related to the fact that the optimization technologies would not be used by companies with sufficient funds to afford them but rather by end-users for whom such an investment would not be profitable.

Haag et al. (2022) showed that many companies are unable to exploit the potential of AI models. Our decision tree helps to provide an initial orientation. However, it is important to efficiently balance the use of AI applications, as not all corporate tasks require AI or XAI solutions and services.

XAI and ethical considerations must be regarded independently; understanding AI decision-making does not mean that the tasks performed are ethical. When analyzing features such as images and videos, it is important to consider whether these tasks are necessary, e.g., for human resource management or loan allocation. In this context, XAI models can only serve as explainable and justified support. Individual human decisions must always be included in such tasks, and human-centric needs, accountability, and decision-making must have high priority. In the health sector, potential applications that are difficult to achieve without AI solutions, such as recognizing biomarkers in computed tomography photos can be exploited.

Theoretical and practical contributions

We contribute to XAI theory by combining literature-based XAI design options with real-world XAI services and develop a decision support framework for academics and practitioners.

Our MBox enhances the understanding of how XAI models can be designed and helps stakeholders to determine which objectives are to be targeted. It also serves as a glossary for the XAI-related vocabulary.

Our research shows how to derive specific MBox and business model archetypes as well as a decision support framework. We use morphological analysis, cluster analysis, and a rule-mining algorithm. According to Osterwalder et al. (2005) and Weking et al. (2020), we build on the three levels of business models: business model elements (MBox), real-world instances (XAI services), and patterns (archetypes). Based on this, we developed a decision support framework to reduce the archetypes' complexity and

[35] http://www.datarobot.com/.

provided a simplified, strategic orientation in the domain of XAI models and their target functions.

Meanwhile, we offer a first market overview and decision support framework to help practitioners to identify the most important XAI design elements. For decision-makers such as managers and data scientists, the decision tree serves as a guide to which XAI design elements are necessary. The decision tree can be used to identify the most appropriate XAI business model and archetype. Based on this, decision-makers can refine their search in the XAI services purchasing process and filter targeted XAI methods or required input data. Even if managers want to program XAI in-house with AI developers, decision-makers can better target a project by narrowing down the development process. Furthermore, the decision tree provides an orientation about what requirements the programming should address and which developers should be engaged. In addition, decision-makers now know which design elements to incorporate and which to dispense.

For regulators and customers affected by the decision regarding an XAI model, our MBox, our archetypical business models, and our decision tree increases AI acceptance; familiarity with XAI design options and business models reduces uncertainty and fear of AI. Moreover, AI developers can use our research for initial guidance on which design options are important for their targeted tasks. XAI service providers can situate their services in the current market and conduct actions to innovate those services. AI service providers who want to expand their business model to XAI get an initial overview, explore the market for their market entry, and identify chances and challenges. For a market entry, the decision tree can be used to select a direction with an archetype and thus align the service with the central design options.

While our MBox provides a comprehensive and complex representation of literature-based XAI design options, our decision tree offers a simplified overview of the dependencies between the most important XAI design dimensions. The seven deduced archetypical business models can be used to benchmark XAI services. In addition, the MBox can be used to develop XAI models or services by selecting one characteristic per dimension to obtain an optimal solution combination iteratively. This can facilitate project work concerning XAI model development as the solution combinations define clear targets. Our research delivers a decision support framework for XAI users, companies, and other XAI stakeholders seeking to adapt and integrate XAI solutions. This is important since many companies have limited knowledge about the potential benefits of AI and XAI solutions in regard to their corporate needs. Therefore, the questions in our decision tree can be answered with little knowledge, reducing the entrance threshold for XAI. In addition, the boxes under the decision classes are recommendations for designers or providers of XAI solutions to consider the most important design elements.

Limitations and further research

One limitation of our study is the subjectivity of our literature review and the classification of the real-world XAI services. To mitigate this limitation, all authors independently reviewed the literature and classified the real-world XAI services. The low number of classified objects due to the limited availability of XAI services is a further constraint that we balanced by integrating XAI services from several industries and application areas. Furthermore, we are unaware of how many customers the XAI services have and to what extent they are satisfied with the services. In addition, it is possible that further XAI services may not fit into one of the identified archetypical business models. Nevertheless, our MBox, archetypical business models, and decision tree can be expanded in further research. Indeed, our decision support framework serves only as a first orientation for practitioners and researchers to reduce entrance thresholds and complexity.

Our MBox, cluster analysis results, archetypical business models, and decision tree provide an extendable basis that further research can build on both quantitatively and qualitatively. Primarily, a next research step can use focus group discussions to evaluate our archetypical business models and our decision tree with practitioners implementing XAI for corporate processes. Moreover, further research can extend the MBox and archetypical business models by investigating the relationships between the characteristics or developing a maturity model (Becker et al., 2009). For example, maturity levels from non-existent to optimized (Becker et al., 2009) can describe XAI models' interpretability, input data types, XAI methods, and value creation. Our MBox offers the possibility to set a detailed research strategy, for example, by focusing exclusively on one dimension of the MBox or selecting one of the three layers. We encourage researchers to conduct more case studies on XAI models and real-world XAI services to further investigate their usefulness and applicability for particular business operations. In addition, in case studies, the surplus in knowledge, explanation, and justification, in contrast to black-box AI models, can be evaluated and discussed to determine whether this protentional surplus is valuable or whether not knowing certain details, such as not understanding the code for specific stakeholders or application areas.

A heat map of XAI research based on our MBox or archetypes can further contribute to theory. Here, a comprehensive literature review can be used to study less-explored but highly relevant research areas. Further research can use a matrix similar to Schoormann et al. (2021) to determine which research topics have been well explored and which research needs deserve more attention. Whether there is actually a need for research must be discussed with practical insights from our archetypal business models providing initial guidance. Taxonomies as proposed by Nickerson et al. (2013) for special XAI models or services (e.g., financial or medical XAI services) can provide a more detailed market overview for practitioners. It can also be

useful to analyze which critical success factors influence XAI use across several sectors (Boynton & Zmud, 1984), identifying which challenges exist and how various real-world applications can be improved or adapted to expand XAI usage.

Conclusions

To address RQ1, which considers the literature-based morphological analysis, we identified 22 scientific publications and grouped them in a classification framework of XAI design options. We built our MBox containing three layers, 13 dimensions, and 51 characteristics. The MBox served as the basis for addressing RQ2 and RQ3, but it also made a theoretical and practical contribution on its own. RQ2 addresses the identification of archetypical business models and decision support for identifying a suitable archetype. We classified 40 real-world XAI services with a broad scope of various application areas and deduced our seven archetypical XAI business models by employing a cluster analysis. Based on our results, we developed a decision support framework in the form of our decision tree, which practically enables XAI stakeholders such as managers, data scientists, and AI developers to select a suitable business model to meet their requirements. To address RQ3, which focuses on the similarities and differences in research and practice, we compared the results of our cluster analysis with those of our MBox. We observed that XAI methods are often used in real-world services that offer recommendations through visualizations, e.g., changes in input data and their influence on the AI model, or graphical representations. Simplified AI models or numerical interpretations of the models that do not provide recommendations are less frequently used. To build and maintain the acceptance of regulatory authorities, users, and customers who are affected by AI, it is important to balance both goals: explanation and decision support. Our MBox, cluster analysis, and decision tree provide a theoretical and practical knowledge base for further theorization and applicable decision support for implementing XAI solutions in business processes and services Table 4.

Appendix 1

Table 4 List of analyzed XAI services

XAI Services	Website
Aignostics	http://www.aignostics.com
Akai Kaeru	https://www.akaikaeru.com
ArthurAI	https://www.arthur.ai/
Beyond Limits	https://www.beyond.ai/beyond-limits-products/
Bolesian	https://bolesian.ai
Clearbox AI Solutions	https://www.clearbox.ai/
Cognino AI	https://www.cognino.ai/
Corpy & Co.	https://corpy.co/
Cycorp	http://www.cyc.com
Dataiku	https://www.dataiku.com/de/
DataRobot	http://www.datarobot.com/
Datricks	https://www.datricks.com
DEEPECHO	https://deepecho.io
Deeploy	https://deeploy.ml
Demystify	https://www.demystify-ai.com
SwiftEnterprise	https://www.digite.com
RISHI-XAI	https://www.digite.com/
DreamQuark	http://www.dreamquark.com/
Dydon	http://www.dydon.net
Fiddler	https://Fiddler.ai/
Google Cloud Explainable AI	https://cloud.google.com/explainable-ai?hl=de
Hacarus	http://hacarus.com/
HMX.ai	https://www.hmx.ai/
Intelligent Artifacts	https://www.intelligent-artifacts.com/
iVCV	https://ivcv.eu/

Table 4 (Continued)

XAI Services	Website
Lagoon	https://www.data-lagoon.com/
Logical Glue	http://www.logicalglue.com
Minerva Intelligence Inc.	https://minervaintelligence.com
neurocat	https://www.neurocat.ai/
Scalnyx	https://www.scalnyx.com
SensViz	https://sensviz.com
Sigma360	https://www.sigma360.ai/
SPIN Analytics	http://spin-analytics.com
Stratyfy	http://www.stratyfy.com
Stride.ai	http://www.stride.ai/
XAI AM	http://www.xai-am.com/
xcoring	https://www.xcoring.ai/
Z Advanced Computing	http://www.zadvancedcomputing.com
Zegami	https://zegami.com/
ZEST AI	https://zest.ai/

Funding Open Access funding enabled and organized by Projekt-DEAL. The research project 'SiNED – Systemdienstleistungen für sichere Stromnetze in Zeiten fortschreitender Energiewende und digitaler Transformation' acknowledges the support of the Lower Saxony Ministry of Science and Culture through the 'Niedersächsisches Vorab' grant programme (grant ZN3563) and of the Energy Research Centre of Lower Saxony.

References

Adadi, A., & Berrada, M. (2018). Peeking inside the black-box: A survey on explainable artificial intelligence (XAI). *IEEE Access, 6*, 52138–52160. https://doi.org/10.1109/ACCESS.2018.2870052

Ahlburg, D. A. (1995). Simple versus complex models: Evaluation, accuracy, and combining. *Mathematical Population Studies, 5*(3), 281–292. https://doi.org/10.1080/08898489509525406

Alamri, R., & Alharbi, B. (2021). Explainable student performance prediction models: A systematic review. *IEEE Access, 9*, 33132–33143. https://doi.org/10.1109/ACCESS.2021.3061368

Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K.-R., & Samek, W. (2015). On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PLoS One, 10*(7), 1–46. https://doi.org/10.1371/journal.pone.0130140

Barocas, S., Selbst, A.D., & Raghavan, M. (2020). The hidden assumptions behind counterfactual explanations and principal reasons. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. https://doi.org/10.1145/3351095.3372830

Barredo Arrieta, A., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., García, S., Gil-López, 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/10.48550/arXiv.1910.10045

Becker, J., Knackstedt, R., & Pöppelbuß, J. (2009). Developing maturity models for IT management. *Business & Information Systems Engineering, 1*, 213–222. https://doi.org/10.1007/s12599-009-0044-5

Bennetot, A., Laurent, J.-L., Chatila, R., & Díaz-Rodríguez, N. (2019). Towards explainable neural-symbolic visual reasoning. *Proceedings of the 28th International Joint Conference on Artificial Intelligence, Macao, China*. https://doi.org/10.48550/arXiv.1909.09065

Boynton, A. C., & Zmud, R. W. (1984). An assessment of critical success factors. *Sloan Management Review, 25*(4), 17–27.

Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Caliskan, A., Bryson, J. J., & Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. *Science, 356*(6334), 183–186. https://doi.org/10.1126/science.aal4230

Carvalho, D. V., Pereira, E. M., & Cardoso, J. S. (2019). Machine learning interpretability: A survey on methods and metrics. *Electronics, 8*(8), 832. https://doi.org/10.3390/electronics8080832

Crupi, R., Castelnovo, A., Regoli, D., & González, B.S. (2021). *Counterfactual explanations as interventions in latent space*. https://doi.org/10.48550/arXiv.2106.07754

Curia, F. (2021). Features and explainable methods for cytokines analysis of dry eye disease in HIV infected patients. *Healthcare Analytics, 1*, #100001. https://doi.org/10.1016/j.health.2021.100001

Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. *Machine Learning*, 1–13. Available at: http://arxiv.org/abs/1702.08608. Accessed 31 May 2022.

Förster, M., Hühn, P., Klier, M., & Kluge, K. (2021). Capturing users' reality: a novel approach to generate. Coherent counterfactual explanations. *Proceedings of the 54th Hawaii International Conference on System Sciences, Maui, USA (virtual).*

Gerlings, J., Shollo, A., & Constantiou, I. (2021). Reviewing the need for explainable artificial intelligence (xAI). *Proceedings of the 54th Hawaiian International Conference on System Sciences, Maui, USA (virtual)*.

Gilpin, L. H., Bau, D., Yuan, B. Z., Bajwa, A., Specter, M., & Kagal, L. (2018). Explaining explanations: an overview of interpretability of machine learning. *Proceedings of the IEEE 5th International Conference on Data Science and Advanced Analytics, Turin, Italy*. https://doi.org/10.1109/DSAA.2018.00018

Goddard, K., Roudsari, A., & Wyatt, J. C. (2012). Automation bias: A systematic review of frequency, effect mediators, and mitigators. *Journal of the American Medical Informatics Association, 19*(1), 121–127. https://doi.org/10.1136/amiajnl-2011-000089

Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., & Pedreschi, D. (2019). A survey of methods for explaining black box models. *ACM Computing Surveys, 51*(5), 1–42. https://doi.org/10.1145/3236009

Haag, F., Hopf, K., Menelau Vasconcelos, P., & Staake, T. (2022). Augmented cross-selling through explainable AI-A case from energy retailing. *Proceedings of the 30th European Conference on Information Systems, Timisoara, Romania*.

Haenlein, M., & Kaplan, A. (2019). A brief history of artificial intelligence: On the past, present, and future of artificial intelligence. *California Management Review, 61*(4), 5–14. https://doi.org/10.1177/0008125619864925

Hakkoum, H., Idri, A., & Abnane, I. (2021). Assessing and comparing interpretability techniques for artificial neural networks breast Cancer classification. *Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization, 9*(6), 587–599. https://doi.org/10.1080/21681163.2021.1901784

Hamm, P., Wittmann, H. F., & Klesel, M. (2021). Explain it to me and I will use it: a proposal on the impact of explainable AI on use behavior. *Proceedings of the 42nd International Conference on Information Systems, Austin, USA*.

Hemmer, P., Schemmer, M., Riefle, L., Rosellen, N., Vössing, M., & Kuehl, N. (2022). Factors that influence the adoption of human-AI collaboration in clinical decision-making. *Proceedings of the 30th European Conference on Information Systems, Timisoara, Romania*.

HLEG-AI. (2019). Ethics guidelines for trustworthy artificial intelligence. *Independent High-Level Expert Group on Artificial Intelligence set up by the European Commission*. Available at: https://www.aepd.es/sites/default/files/2019-12/ai-ethics-guidelines.pdf. Accessed 31 May 2022.

Ivaturi, P., Gadaleta, M., Pandey, A. C., Pazzani, M., Steinhubl, S. R., & Quer, G. (2021). A comprehensive explanation framework for biomedical time series classification. *IEEE Journal of Biomedical and Health Informatics, 25*(7), 2398–2408. https://doi.org/10.1109/JBHI.2021.3060997

Kamiński, B., Jakubczyk, M., & Szufel, P. (2018). A framework for sensitivity analysis of decision trees. *Central European Journal of Operations Research, 26*, 135–159. https://doi.org/10.1007/s10100-017-0479-6

Kaplan, A., & Haenlein, M. (2019). Siri, Siri, in my hand: Who's the fairest in the land? On the interpretations, illustrations, and implications of artificial intelligence. *Business Horizons, 62*(1), 15–25. https://doi.org/10.1016/j.bushor.2018.08.004

Kaufman, L., & Rousseeuw, P. J. (1990). *Finding groups in data*. Wiley & Sons.

Kibria, M. G., Nguyen, K., Villardi, G. P., Zhao, O., Ishizu, K., & Kojima, F. (2018). Big data analytics, machine learning, and artificial intelligence in next-generation wireless networks. *IEEE Access, 6*, 32328–32338. https://doi.org/10.1109/ACCESS.2018.2837692

Kim, T. W. (2018). *Explainable artificial intelligence (XAI), the goodness criteria and the grasp-ability test*. ArXiv,1–7. Available at: http://arxiv.org/abs/1810.09598. Accessed 31 May 2022.

Koziol, C., & Weitz, S. (2021). Does model complexity improve pricing accuracy? The case of CoCos. *Review of Derivatives Research, 24*, 261–284. https://doi.org/10.1007/s11147-021-09178-4

Kridel, D., Dineen, J., Dolk, D., & Castillo, D. (2020). Model interpretation and Explainability: towards creating transparency in prediction models. *Proceedings of the 53th Hawaii International Conference on System Sciences, Maui, USA*.

Kundisch, D., Muntermann, J., Oberländer, A.M., Rau, D., Röglinger, M., Schoormann, T., & Szopinski, D. (2021). An update for taxonomy designers. *Business & Information Systems Engineering*. Online first. https://doi.org/10.1007/s12599-021-00723-x.

Li, X. -H., Cao, C. C., Shi, Y., Bai, W., Gao, H., Qiu, L., Wang, C., Gao,Y., Zhang, S., Xue, X., & Chen, L. (2020). A survey of data-driven and knowledge-aware explainable AI. *IEEE Transactions on Knowledge and Data Engineering, 34*(1), 29–49. https://doi.org/10.1109/TKDE.2020.2983930

Linardatos, P., Papastefanopoulos, V., & Kotsiantis, S. (2021). Explainable AI: A review of machine learning interpretability methods. *Entropy, 23*(1), 1–45. https://doi.org/10.3390/e23010018

Lipton, Z. C. (2018). The mythos of model interpretability. *Communications of the ACM, 61*(10), 36–43. https://doi.org/10.1145/3233231

Lockey, S., Gillespie, N., Holm, D., & Someh, I. A. (2021). A review of trust in artificial intelligence: challenges, vulnerabilities and future directions. *Proceedings of the 54th Hawaii International Conference on System Sciences, Maui, USA* (virtual).

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Proceedings of the 31th Conference on Neural Information Processing Systems, Long Beach, USA*.

Markus, A. F., Kors, J. A., & Rijnbeek, P. R. (2021). The role of explainability in creating trustworthy artificial intelligence for health care: A comprehensive survey of the terminology, design choices, and evaluation strategies. *Journal of Biomedical Informatics, 113*, 1–11. https://doi.org/10.1016/j.jbi.2020.103655

McKinney, S. M., Sieniek, M., Godbole, V., Godwin, J., Antropova, N., Ashrafian, H., Back, T., Chesus, M., Corrado, G. S., Darzi, A., Etemadi, M., Garcia-Vicente, F., Gilbert, F. J., Halling-Brown, M., Hassabis, D., Jansen, S., Karthikesalingam, A., Kelly, C. J., King, D., Ledsam, J. R., Melnick, D., Mostofi, H., Peng, L., Reicher, J. J., Romera-Paredes, B., Sidebottom, R., Suleyman, M., Tse, D., Young, K. C., De Fauw, J., & Shetty, S. (2020). International evaluation of an AI system for breast cancer screening. *Nature, 577*(7788), 89–94. https://doi.org/10.1038/s41586-019-1799-6

Meister, S., Wermes, M., Stüve, J., & Groves, R. M. (2021). Investigations on explainable artificial intelligence methods for the deep learning classification of fibre layup defect in the automated composite manufacturing. *Composites Part B: Engineering, 224*, #109160. https://doi.org/10.1016/j.compositesb.2021.109160

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

Mohseni, S., Zarei, N., & Ragan, E. D. (2021). A multidisciplinary survey and framework for design and evaluation of explainable AI systems. *ACM Transactions on Interactive Intelligent Systems, 11*(3–4), 1–45. https://doi.org/10.1145/3387166

Nickerson, R. C., Varshney, U., & Muntermann, J. (2013). A method for taxonomy development and its application in information systems. *European Journal of Information Systems, 22*(3), 336–359. https://doi.org/10.1057/ejis.2012.26

Omdia. (2021). *Revenues from the artificial intelligence (AI) software market worldwide from 2018 to 2025*. Available at: https://www.

statista.com/statistics/607716/worldwide-artificial-intelligence-market-revenues/. Accessed 31 May 2022.

Osterwalder, A., Pigneur, Y., & Tucci, C. L. (2005). Clarifying business models: Origins, present, and future of the concept. *Communications of the Association for Information Systems, 16*(1), 1–25. https://doi.org/10.17705/1CAIS.01601

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Müller, A., Nothman, J., Louppe, G., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É. (2011). Scikit-learn: Machine learning in python. *Journal of machine learning research, 12*, 2825–2830.https://doi.org/10.48550/arXiv.1201.0490

Polzer, A. K., Fleiß, J., Ebner, T., Kainz, P., Koeth, C., & Thalmann, S. (2022). Validation of AI-based information systems for sensitive use cases: using an XAI approach in pharmaceutical engineering. *Proceedings of the 55th Hawaii International Conference on System Sciences, Maui, USA (virtual)*.

Punj, G., & Stewart, D. W. (1983). Cluster analysis in marketing research: Review and suggestions for application. *Journal of Marketing Research, 20*(2), 134–148. https://doi.org/10.2307/3151680

Rai, A. (2020). Explainable AI: From black box to glass box. *Journal of the Academy of Marketing Science, 48*(1), 137–141. https://doi.org/10.1007/s11747-019-00710-5

Ritchey, T. (2011). Modeling alternative futures with general morphological analysis. *World Futures Review, 3*(1), 83–94. https://doi.org/10.1177/194675671100300105

Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics, 20*, 53–65. https://doi.org/10.1016/0377-0427(87)90125-7

Ruck, D. W., Rogers, S. K., & Kabrisky, M. (1990). Feature selection using a multilayer perceptron. *Journal of Neural Network Computing, 2*, 40–48.

Saputra, D. M., Saputra, D., & Oswari, L. D. (2020). Effect of distance metrics in determining k-value in k-means clustering using elbow and silhouette method. *Proceedings of the Sriwijaya International Conference on Information Technology and its Applications, Palembang, Indonesia*. https://doi.org/10.2991/aisr.k.200424.051

Schoormann, T., Strobel, G., Möller, F., & Petrik, D. (2021). Achieving sustainability with artificial intelligence-a survey of information systems research. *Proceedings of the 42nd International Conference on Information Systems, Austin, USA*.

Seppälä, A., Birkstedt, T., & Mäntymäki, M. (2021). From ethical AI principles to governed AI. *Proceedings of the 42nd International Conference on Information Systems, Austin, USA*.

Setzu, M., Guidotti, R., Monreale, A., Turini, F., Pedreschi, D., & Giannotti, F. (2021). GLocalX - from local to global explanations of black box AI models. *Artificial Intelligence, 294*, #103457. https://doi.org/10.1016/j.artint.2021.103457

Sipior, J. C., Lombardi, D. R., & Gabryelczyk, R. (2021). AI recruiting tools at ShipIt2Me.com. *Communications of the Association for Information Systems, 48*, 443–455. https://doi.org/10.17705/1CAIS.04839

Statista. (2022). *Size of explainable artificial intelligence (AI) market worldwide from 2020 to 2030*. Available at: https://www.statista.com/statistics/1256246/worldwide-explainable-ai-market-revenues/. Accessed 31 May 2022.

Stepin, I., Alonso, J. M., Catala, A., & Pereira-Farina, M. (2021). A survey of contrastive and counterfactual explanation generation methods for explainable artificial intelligence. *IEEE Access, 9*, 11974–12001. https://doi.org/10.1109/ACCESS.2021.3051315

Stroppiana Tabankov, S., & Möhlmann, M. (2021). Artificial intelligence for in-flight services: how the lufthansa group managed explainability and accuracy concerns. *Proceedings of the 42nd International Conference on Information Systems, Austin, USA*.

Templier, M., & Paré, G. (2015). A framework for guiding and evaluating literature reviews. *Communications of the Association for Information Systems, 37*(1), 112–137. https://doi.org/10.17705/1CAIS.03706

Thiebes, S., Lins, S., & Sunyaev, A. (2021). Trustworthy artificial intelligence. *Electronic Markets, 31*(2), 447–464. https://doi.org/10.1007/s12525-020-00441-4

Vilone, G., & Longo, L. (2021). Notions of explainability and evaluation approaches for explainable artificial intelligence. *Information Fusion, 76*, 89–106. https://doi.org/10.1016/j.inffus.2021.05.009

vom Brocke, J., Simons, A., Riemer, K., Niehaves, B., Plattfaut, R., & Cleven, A. (2015). Standing on the shoulders of giants: Challenges and recommendations of literature search in information systems research. *Communications of the Association for Information Systems, 37*(1), 205–224. https://doi.org/10.17705/1CAIS.03709

Wambsganss, T., Engel, C., & Fromm, H. (2021). Improving explainability and accuracy through feature engineering: a taxonomy of features in NLP-based machine learning. *Proceedings of the 42nd International Conference on Information Systems, Austin, USA*.

Wang, H., Li, C., Gu, B., & Min, W. (2019). Does AI-based credit scoring improve financial inclusion? Evidence from online payday lending. *Proceedings of the 40th International Conference on Information Systems, Munich, Germany*.

Wastensteiner, J., Weiss, T. M., Haag, F., & Hopf, K. (2021). Explainable AI for tailored electricity consumption feedback - an experimental evaluation of visualizations. *Proceedings of the 29th European Conference on Information Systems, Marrakesh, Morocco* (virtual).

Watson, R. T., & Webster, J. (2020). Analyzing the past to prepare for the future: Writing a literature review a roadmap for release 2.0. *Journal of Decision Systems, 29*(3), 129–147. https://doi.org/10.1080/12460125.2020.1798591

Webster, J., & Watson, R. T. (2002). Analyzing the past to prepare for the future: Writing a literature review. *MIS Quarterly, 26*(2), xiii–xxiii.

Weking, J., Mandalenakis, M., Hein, A., Hermes, S., Böhm, M., & Krcmar, H. (2020). The impact of blockchain technology on business models – A taxonomy and archetypal patterns. *Electronic Markets, 30*(2), 285–305. https://doi.org/10.1007/s12525-019-00386-3

Xie, J., Chai, Y., & Liu, X. (2022). An interpretable deep learning approach to understand health. Misinformation transmission on YouTube. *Proceedings of the Hawaii 55th International Conference on System Sciences, Maui, USA* (virtual).

Zhang, K., Xu, P., Gao, T., & Zhang, J. (2021). A trustworthy framework of artificial intelligence for power grid dispatching systems. *Proceedings of the IEEE International Conference on Digital Twins and Parallel Intelligence, Beijing, China*. https://doi.org/10.1109/DTPI52967.2021.9540198

Zschech, P., Weinzierl, S., Hambauer, N., Zilker, S., & Kraus, M. (2022). GAM(e) changer or not? An evaluation of interpretable machine learning models based on additive model constraints. *Proceedings of the 30th European Conference on Information Systems, Timisoara, Romania*.

Zwicky, F. (1967). The morphological approach to discovery, invention, research and construction. In F. Zwicky & A. G. Wilson (Eds.), *New methods of thought and procedure*. Springer.