Explainable AI: a narrative review at the crossroad of Knowledge Discovery, Knowledge Representation and Representation Learning

Ikram Chraibi Kaadoud[1*], **Lina Fahed**[1], **Philippe Lenca**[1]
[1]IMT Atlantique, Lab-STICC, UMR CNRS 6285, F-29238 Brest, France
{ikram.chraibi-kaadoud, lina.fahed, philippe.lenca}@imt-atlantique.fr

Abstract

EXplainable Artificial Intelligence (XAI) has recently become a very active domain, mainly due to the extensive development of black-box models such as neural networks. Recent XAI objectives have been defined in the state-of-the-art, for which specific approaches have been proposed. Implicit links can be found between XAI and other domains, especially related to knowledge and neural networks. We here aim to highlight these implicit links. We present a narrative review of research works in two domains: (i) Knowledge domain with focus on Knowledge Discovery and Representation, and (ii) Representation Learning. We discuss the similarity and joining points between these domains and XAI. We conclude that, in order to make black-boxes more transparent, XAI approaches should be more inspired and take advantage of past and recent works in Knowledge and Representation Learning domains. Through this paper, we offer an entry point to the domain of XAI for both multidisciplinary researchers and specialists in AI, as well for AI knowledgeable users.

Keywords: XAI, Knowledge Discovery, Knowledge representation, Representation learning, State representation learning, Manifold representation learning, Multi-view representation learning, Network representation learning

1 Introduction: XAI

During the last few years, eXplainable Artificial Intelligence (XAI), has become a very active domain[1] facing the high development of black-box models, such as neural networks [Guidotti *et al.*, 2018]. A new generation of XAI approaches have been proposed, for which several new concepts and terms are specific to application domains, data types or modeling. Application domains of XAI are multiple: machine learning, robotics, multi-agent systems, computer vision, Knowledge Representation and Reasoning, *etc.*

[Barredo Arrieta *et al.*, 2020] defined *"Given an audience, an explainable Artificial Intelligence is the one that produces details or reasons to make its functioning clear or easy to understand"*. Indeed, XAI aims to make Artificial Intelligence (AI) models more intelligible and accessible or to directly design explainable models and results [Buchanan and Shortliffe, 1984; Guidotti *et al.*, 2018; Barredo Arrieta *et al.*, 2020]. When the first case arises, XAI provides an *explanation* of the internal mechanisms and/or the reasons behind the AI model behavior *i.e.* its functioning and performance: an explanation is thus an interface between the AI model to explain and the target audience [Gunning, 2017]. We define an explanation as an information in a semantically complete format, which is self-sufficient and chosen according to the target audience regarding its knowledge, its expectations and the context. Hence, the purpose of an explanation is to clarify the cause, context and consequences of described facts through a set of statements or information [Van Fraassen, 1988].

It is important to underline that an explanation by its very nature is contextual: it is specific to a given target audience and also to a given context [Walton, 2004]. This makes XAI more challenging as automatic context understanding is still a very challenging task [Brézillon, 1999; Lim *et al.*, 2009; Augusto *et al.*, 2017; Hollister *et al.*, 2017] and no unified way for modelling context in intelligent environments has yet been proposed in the literature [Brenon *et al.*, 2018]. We emphasize that the context (*i.e* users context, goal context, *etc.*) is important to take into account in XAI. However, this point is not the focus of the paper.

In the state-of-the-art, an explanation can take different formats (*e.g.* visual, natural language, features relevance explanations, *etc.*) and combine several representations of the same information [Barredo Arrieta *et al.*, 2020]. Two main XAI techniques are proposed: (i) Ante-hoc techniques which consist in optimizing an already transparent AI model (*e.g.* linear regression, decision trees, *etc.*) by adding constraints or features in order to increase transparency through metrics, data visualisation, *etc.* (ii) Post-hoc techniques that aim to explain already built black-box AI models (mainly deep neural networks). Among famous Post-hoc techniques: LIME [Ribeiro *et al.*, 2016], SHAP [Lundberg and Lee,

*Contact Author

[1]We remind that Artificial Intelligence models with explanation goals have been questioned and investigated a long time ago such as in [Shortliffe, 1974]. However, the term XAI has been recently proposed.

2017], visual explanations, saliency mapping, *etc.* XAI has recently been covered by several reviews that reveal its complex and intrinsically multidisciplinary aspects from a technical, user or Human-Interaction viewpoint [Guidotti *et al.*, 2018; Gilpin *et al.*, 2018; Barredo Arrieta *et al.*, 2020; Vilone and Longo, 2020]. As examples, we can note technical-based reviews, as those related to reinforcement learning [Puiutta and Veith, 2020; Heuillet *et al.*, 2021], data-based reviews as those related to time series [Schlegel *et al.*, 2019; Rojat *et al.*, 2021] and application-based reviews related to healthcare [Adadi and Berrada, 2020] and banking [Burgt, 2020]. Other reviews are inspired by social science, human psychology, sociology or cognitive sciences [Miller, 2019; Capone and Bertolaso, 2020] in order to build ethical and fair models [Barredo Arrieta *et al.*, 2020].

One key issue that have not been discussed in the above cited reviews and that we would like to highlight, is the importance of knowledge in XAI. As an interface between an AI and a target audience, an explanation can be considered as an interpreter between the AI knowledge and the human target audience knowledge. Since knowledge domain is historical in AI, this raises in turn important questions about the impact of domains such as Knowledge Discovery and Representation on XAI. Furthermore, regarding black-box models and especially neural networks, it is important to mention that in recent papers, concepts like representation learning, knowledge/latent/hidden/abstract representation, latent space, *etc.* have been studied in order to tackle issues such as dimensionality, running time, algorithmic complexity, *etc.* However, to the best of our knowledge, no explicit relation has been defined between these concepts and XAI. We consider that as these concepts are increasingly recurrent in the literature, with no consensual definitions across fields, it becomes, in turn, more difficult to apprehend the XAI domain.

To address this issue, we propose a narrative review that, contrary to the above cited literature reviews, does not review XAI techniques. Our paper is a narrative review across several domains: a literature-based review that synthesizes technical research works related to domains that implicitly inspire XAI works. Our goal is to bring original insights, formulate new research questions and highlight promising future directions of XAI. More precisely, in this narrative review, we aim to address three questions. **First,** to centralize and clarify concepts recurrently used in AI domains but not always clear for XAI specialists. **Second**, to bring a new light to XAI by making explicit the links between XAI and two other domains: (i) Knowledge domain including Knowledge Discovery Process (KDP) and Knowledge Representation (KR), and (ii) Representation Learning (RL) more associated to deep learning domain. **Third**, to offer an entry point to the XAI domain for multidisciplinary or specialists in these domains.

Actually, these domains are often perceived as disconnected as most of the research is currently concentrated on only one of them [Sallinger *et al.*, 2020]. Despite this, we believe that it is important to enhance the links and the implicit relations that can be found between them. We thus consider that XAI has been indirectly inspired by these domains.

Figure 1 shows our vision as a schematic representation of XAI domain and both KDP, KR and RL domains. Table 1 lists the acronyms used. The paper is organized as follows: definitions are presented in section 2, KDP and KR in section 3, and RL in section 4. At the end of both last sections, we discuss the relation between the highlighted points, related to KDP, KR, RL and XAI. Finally, in section 5, we discuss future directions and perspectives related to XAI.

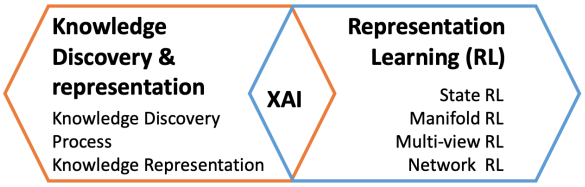

Figure 1: A schematic representation of XAI and its positioning at the crossroads of other domains.

Acronym	Research domain
XAI	eXplainable Artificial Intelligence
KDP	Knowledge Discovery process
KR	Knowledge Representation
RL	Representation Learning
SRL	State Representation Learning

Table 1: Acronyms of research domains discussed in this paper.

2 Definitions

This section is dedicated to the definition of several concepts related to Knowledge and Representation learning domains. Several definitions are inspired from state-of-the-art works.

Definition 2.1 The raw material that represents the input of an algorithm is called **data**. Data can be noisy, partial/complete, un/structured and of different types [Grazzini and Pantisano, 2015; Malhotra and Nair, 2015].

Definition 2.2 A data set is a collection of data that describes real-word **objects** (such as cars, documents, animal, *etc.*) through multiple properties called **features** [Bishop, 2006].

Definition 2.3 Once data is analyzed and correlated, it represents **information**. Information can be reproduced from data and its importance depends on the context it is generated from/for [Grazzini and Pantisano, 2015; Malhotra and Nair, 2015].

Definition 2.4 **Knowledge** is a set of information that is assessed by a human, *i.e.* human adds a value and semantics according to his/her own background and context [Grazzini and Pantisano, 2015; Malhotra and Nair, 2015].

Definition 2.5 In the data mining domain, a "***pattern*** *is an expression in some language describing a subset of the data or a model applicable to the subset*" [Fayyad *et al.*, 1996]. Hence, **Pattern extraction** designates the process of finding structures in data, fitting a model to data, or finding a high-level description of a data set.

Many data modeling approaches have been proposed in the state-of-the-art. We can cite reinforcement learning, graph-based approaches, neural networks, *etc.* We now define some important concepts related to these approaches.

Definition 2.6 Reinforcement learning is an approach in which an intelligent **agent** interacts with its environment through trial-and-errors actions in order to reach a goal. Each action leads to a modification of the **state** of the agent and the environment and the increase or decrease of a cumulative **reward** value. Actions are chosen according to a strategy that is called a **policy** [Barto and Sutton, 1995].

Definition 2.7 A **Manifold** is a topological structure of n-dimensions. For example, a one-dimensional manifold is a curve, a two-dimensional manifold is a surface, a three-dimensional manifold is a sphere.

Definition 2.8 A **network** is a collection of discrete objects called **nodes**, which are connected through **links**: it can be viewed as a graph with **vertices** and **edges**, both with attributes/weights or not [Fletcher *et al.*, 1991].

Definition 2.9 Neural networks are machine learning models with several architectures, that are usually structured by one or several layers (input, hidden and output). Each layer is composed of one or several computational **units** called artificial neurons - conceptually derived from biological neurons [McCulloch and Pitts, 1943; Abraham, 2005]. Computational units can also be a Long Short Term Memory (well known also as LSTM) [Hochreiter and Schmidhuber, 1997] or Gated recurrent units [Cho *et al.*, 2014]. A **deep neural network** have many hidden layers, units, and edges with weights. Units of layer n can be all or partially connected to units of layer $n+1$. Due to this inner complexity, deep neural networks are a typical example of **black-boxes**.

Definition 2.10 In neural networks, an **activation pattern** refers to units activation values of one of the layers. An activation pattern is a numerical **vector** of the size of the layer it is associated with. A **hidden pattern** refers to the activation pattern of a hidden layer.

In the literature of neural networks, concepts like latent space and latent representation have been developed and widely used. However, to the best of our knowledge, no complete definitions have been clearly proposed for such concepts. Due to the importance of both concepts in the rest of this paper, we choose to formulate their definition next.

Definition 2.11 Latent space refers to the abstract multi-dimensional space associated to each layer of a neural network where the representation of the learned data is implicitly built. Latent space contains the meaningful internal features (definition 2.2) representations of learned data, which makes it not directly interpretable. In a deep neural network (definition 2.9), each hidden layer, whether it has the same number of units or not, has its own latent space. It is thus possible to extract several implicit representations from this network. The latent space can be used to achieve a data dimensionality reduction, when the hidden layer is smaller than the input layer. This is the case for example with autoencoders and variational autoencoders [Kingma and Welling, 2014], models that can reduce high-dimensional inputs into efficient and representative low-dimensional representations [Roberts *et al.*, 2018b].

Definition 2.12 Latent or hidden representation refers to the data representation implicitly encoded by a neural network during the learning task and thus is hidden-layer-dependant [Bengio *et al.*, 2013]. It is a machine-readable data representation that contains features of the original data that have been learned by associated hidden layer. One key property of latent space (definition 2.11) is that real-world objects (definition 2.2) that are semantically close (*e.g.* cars of different brands), will end up grouped together in one latent space: their respective hidden representation in the corresponding layer, will be close to each other compared to other objects that are not semantically close (*e.g.* cats) [Roberts *et al.*, 2018a]. Thus, a latent representation is useful for pattern analysis (definition 2.5) and for similarity detection between objects (definition 2.2) using clustering methods.

3 Knowledge: discovery and representation

We now present two active research domains: KDP (section 3.1) and KR (section 3.2). Then, we discuss the relation between them and XAI in section 3.3.

3.1 Knowledge Discovery Process (KDP)

KDP is a human-centered domain that seeks useful knowledge (definition 2.4) through an iterative and interactive process that involves humans [Lenca, 2002; Cios *et al.*, 2007]. As the domains KDP, data mining, and Knowledge Discovery in Databases (referred to as KDD) are often used in a confused way, we consider that it is important to present a clarification about them, as follows:

- According to [Cios *et al.*, 2007], KDP and KDD designate the same process. However, KDP can be generalized to non-databases sources of data, while KDD emphasizes databases as a primary source of data.
- KDP and data mining are related to each other as well as to other domains like machine learning and statistics, but are clearly distinct. Indeed, according to [Fayyad *et al.*, 1996] and [Cios *et al.*, 2007], KDP is the global process of discovering useful knowledge from data, whereas data mining is a particular step within the KDP process that consists in applying algorithms to extract patterns (definition 2.5) or to build a model that fits the data.

There is no consensus about the steps of a KDP: nine steps in [Fayyad *et al.*, 1996], eight steps in [Anand and Büchner, 1998], six steps in [Wirth, 2000; Cios *et al.*, 2007] and five steps in [Cabena *et al.*, 1998]. However, we emphasize that globally KDP consists of three common main steps:

1. A pre-processing step for data collection or generation, data preparation, cleaning, curing, *etc.*
2. A data processing step where several techniques from statistics/machine learning/data mining, *etc.* communities can be used.
3. A post-processing step for visualisation, evaluation and validation.

At each step, the extracted information (definition 2.3) is usually evaluated by the human, given the context, to form knowledge[2] (definition 2.4). Thus, the target audience of the KDP is the human: application domain experts and decision makers. In addition, it is important to underline that two mains goals of KDP are usually defined [Fayyad *et al.*, 1996]: (i) verification of a user hypothesis, and (ii) discovery of valid and useful new knowledge that is understandable with respect to the data (definition 2.1) from which it is derived. These goals are thoroughly discussed in section 3.3.

3.2 Knowledge Representation (KR)

KR is a crucial question in AI [Malhotra and Nair, 2015]. Also known as "Knowledge Representation and Reasoning", KR aims at finding ways to efficiently structure specific domain knowledge for automated reasoning. In this way, intelligent machines can learn, draw inferences, make decision and answer questions related to this knowledge [Davis *et al.*, 1993; Shapiro, 2006; Davis, 2015]. Thus, seen in such a way, KR can be considered as a machine-oriented domain. The purpose of KR is neither about storing data, nor making actions but it is about allowing *"thinking by reasoning"* [Davis *et al.*, 1993]. Consequently, KR has been a key component for the conception of intelligent knowledge-based systems.

KR is also, according to [Malhotra and Nair, 2015], closely related to the Knowledge retrieval in the shape of ontologies (concepts for representing, storing and accessing knowledge [Guarino *et al.*, 2009]). KR techniques have also been widely developed and applied to semantic web [Hagedorn *et al.*, 2020], semantic networks [Malhotra and Nair, 2015], text interpretation and cognitive robotics [Davis, 2015]. In addition, from a user viewpoint, KR is important during the development of software systems in order to perform particular tasks, as well as for broader community of cognitive science whose goal is to constitute and organize knowledge from humans and machine perspectives [Das, 2003].

Knowledge Representation Learning (KRL)

KRL is the process of making AI algorithms model and learn a structured representation of domain-specific knowledge. As a consequence, concepts, relations between them and their representations can be encoded in a low-dimensional semantic space [Lin *et al.*, 2018]. For example, when knowledge is represented as a graph, the KRL process allows graph embedding and preserves semantic similarities [Xie *et al.*, 2018]. Notice that the development of deep learning algorithms and their performance on distributed representations (*i.e.* representations that describe features of the same data across layers) that reduce the computational complexity has contributed to the emergence of several KRL applications such as recommendation system [Zhang *et al.*, 2016], language modeling [Ahn *et al.*, 2016] and question answering [Yin *et al.*, 2016]. We consider that KR has recently become a more central domain in AI, and by extension in XAI. This is mainly due to the development of Representation Learning in neural networks (introduced in section 4).

[2]Notice that recent approaches like AutoML tend to perform all these steps automatically without user intervention [He *et al.*, 2021]

3.3 Discussion: relation between KDP, KR and XAI

We now discuss and highlight several links and common points between KDP, KR and XAI. As mentioned in sections 3.1 and 3.2, KDP is a human-centered domain, whereas KR is a machine-oriented domain. However, both domains are complementary: in KDP, the main question is "How to efficiently discover new or retrieve existing knowledge?", whereas in KR the tackled question is "How to represent the knowledge efficiently to be able to *reason* on it?".

It is important to highlight that both KDP and KR questions are also addressed and are crucial in XAI. Recall that the objective of XAI is to make the reasons behind AI behavior simple and accessible to a target audience regarding a given task and context. We consider that this XAI objective can be viewed and divided into two sub-objectives: (i) to discover the reasons behind AI behavior - which is the same as in a KDP problem -, (ii) to represent these reasons in a way that is intelligible for the human target audience, but also sometimes for an artificial one - which is the same as for a KR problem.

Let us first detail the links between KDP and XAI. In XAI, for black-boxes like deep neural networks [Guidotti *et al.*, 2018; Gilpin *et al.*, 2018], technical approaches are used to search the behavior of AI and make it explainable by providing an explanation that can take several forms and be multimodal [Barredo Arrieta *et al.*, 2020]. Explaining an AI model is therefore very inspired by KDP. The particular point is that in XAI, the input data (definition 2.1) is related to the black-box AI model. This input data can be of several types, *e.g.* activation patterns of hidden layers (definition 2.10), features or representations, and require the same techniques as in KDP.

Figure 2 represents a schematic representation of the transformation of data into knowledge, in KDP and XAI domains. It clarifies the similarities between both domains regarding the human intervention, and the role of the technical part, *i.e.* data mining and explainable methods.

Let us now go into deep details about knowledge representation in XAI. Two cases can be highlighted according to target audience: (i) human who uses the knowledge representation to reason and understand the situation, *e.g.* the decision maker and the application domain expert, depending on their expertise, role and goals, (ii) another AI system for which the input data is provided from a complex AI architecture.

Let us take an example in the domain of computer vision and especially classification using deep neural networks. Researchers have proposed approaches that exploit different AI algorithms and their latent representation (definition 2.12) as an input to the neural networks. The objective of such approaches is to perform both classification and explainability tasks through saliency masks applied to images and text generation [LeCun *et al.*, 2015]. This is one strategy among multiple others for the representation of knowledge in order to favor the explainability of the behavior of the initial AI model.

In addition, notice that KRL has been basically associated with deep learning algorithms, especially with techniques like graph representation learning [Hamilton, 2020] and concept learning [Dolgikh, 2018], which are both studied in the XAI domain [Xu *et al.*, 2018; Fazi, 2020].

Finally, it is important to highlight the importance of the target audience in both KDP/KR and XAI domains. Actually, the role of the target audience is decisive: knowledge is usually retrieved and shaped in order to answer a question of a target audience related to a given task and context such as verifying an hypothesis, inference and decision making. Knowledge representation and content in both KDP/KR and XAI domains are thus target and context dependant.

As a conclusion, XAI is closely related to both KDP and KR, and future works in XAI should take advantage of recent works in both domains, as well as older works.

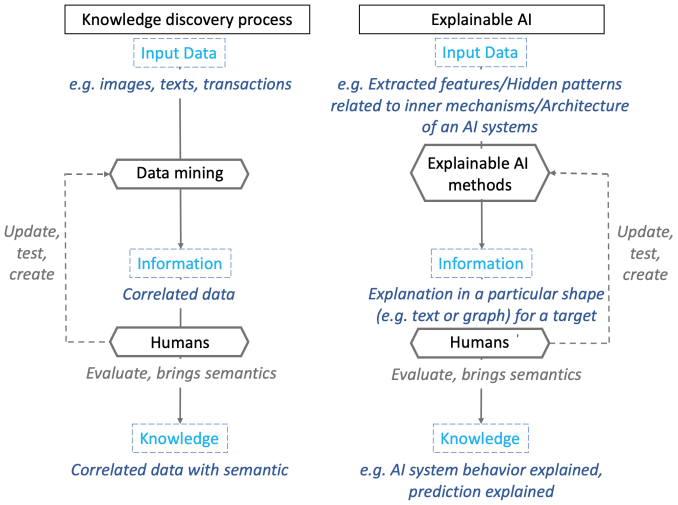

Figure 2: Two schematic ways for data transformation into knowledge: On the left, within a Knowledge Discovery process, and on the right within a XAI process.

4 Representation Learning (RL)

We now present the RL domain, its importance in deep neural networks, and RL sub-domains that are recurrent and popular.

4.1 RL introduction and definition

RL has been discussed as a key challenge related to different machine learning domains [Dietterich *et al.*, 2008] especially to neural networks. As first demonstrated by [Rumelhart *et al.*, 1986], in neural networks, back-propagation algorithms can generate useful internal representations of data in hidden layers. Since then, different approaches have been proposed in order to learn, analyze and visualize latent data representations [Gilpin *et al.*, 2018; Guidotti *et al.*, 2018]. Thus, RL has become an active research domain for which the objective is to study of latent representations in order to improve deep neural network efficiency [Bengio *et al.*, 2013].

RL - and synonyms like Data RL or Feature Learning [Zhong *et al.*, 2016] - focuses on *"learning representations of the data that make it easier to extract useful information when building classifiers or other predictors"* [Bengio *et al.*, 2013]. In other words, RL is designed to learn abstract features that characterize data [Lesort *et al.*, 2018].

RL algorithms can be classified into two categories: global and local RL algorithms. While the first ones tend to preserve the data global information in the learned feature space, the second ones focus more on preserving local similarity between data during learning the new representations [Zhong *et al.*, 2016]. Representations are not task-specific but are useful to machine learning algorithms to solve tasks, as well as to humans to comprehend the behavior of these last algorithms [Bengio *et al.*, 2013]. One of the reasons that makes RL popular is that representations express priors about the data. The expressed priors can vary within a single learning algorithm. Consequently, the characteristic of the priors variations leads to different RL approaches, that we classify into two categories: problems-oriented RL and data-oriented RL.

In the following section, we first present the concept of hierarchical representation in deep neural networks, a key property of RL. Then we present examples of particular cases of RL that are problems-oriented and data-oriented.

4.2 Hierarchical representations in deep neural networks

One key property of the RL domain in deep neural networks is the ability to provide both high level features and low level features for the same learned data. Recall that a deep neural network will encode a latent representation at each hidden layer (definitions 2.9, 2.12). Since the layer n units can be all or partially connected to the layer $n+1$ units, each layer uses the previous layer as input. If the previous layer is a hidden layer, then the input is already a latent representation, *i.e.* an abstract feature representation that characterizes the data. Thus, each layer extracts an abstract feature representation of the previous layer. As a result, a deep neural network learns multiple levels of abstraction and implicitly encodes a **hierarchy of latent and abstract representations** that are built progressively, layer by layer. The layers that are close to the input layer will encode a low-level feature representation, whereas those deeper inside the architecture will encode a high level feature representation. In other words, the closer the considered layer is to the output layer, the more the representation is abstract [Bengio *et al.*, 2013; Zhong *et al.*, 2016; Lesort *et al.*, 2018], as represented in Figure 3.

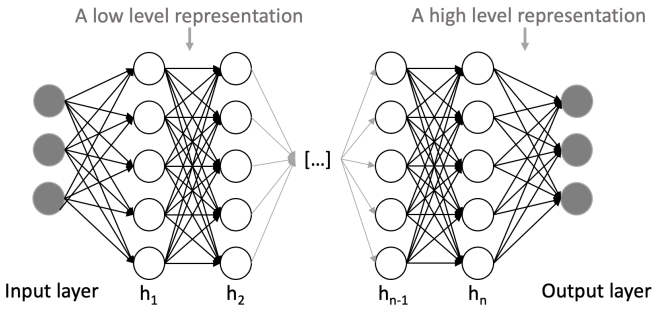

Figure 3: Illustrative and schematic representation of the position of a low level representation and a high level representation in a deep neural network. h_x refers to the x^{th} hidden layer in the network.

It has also been shown that, in deep learning algorithms, hidden representations tend to keep dominant information and propagate them across hidden layers, regardless the width

or depth increase of the deep neural networks [Nguyen *et al.*, 2021]. This characteristic of RL is also a key one for XAI: by extracting and comparing the low-level and the high-level representations of a deep architecture, we consider that it is possible to explicit the inner mechanism of the architecture by observing the differences between the representations. This will be discussed further in section 5.

4.3 Problems-oriented RL approaches

Recall that the objective of RL algorithms is to learn abstract features that characterize data. This objective can be challenging according the issues that one could face such as high dimensionnality of data or RL application to another AI paradigm like reinforcement learning (definition 2.6). In the following sub-sections, we describe two RL sub-domains: Manifold RL and State RL, that have recently shown great performances in deep learning and that deal with our core questions. The links with XAI are also briefly discussed.

Manifold RL

Manifold RL is particularly suited for dealing with high-dimensional data sets that are very difficult to visualize and less intuitive. However, within such data sets, data can locally belong to a subset that can be represented by a manifold. As stated in definition 2.7, a manifold is a topological structure of n-dimensions. Thus, Manifold RL corresponds to the learning of complex data representation in several dimensions while preserving the topological properties of the considered manifold. We consider Manifold RL as a non-linear dimensionality reduction approach, that can help to discover similarities in data for which dimensions have been reduced [Cayton, 2005; Bengio, 2009; Zhang *et al.*, 2011]. The Manifold RL domain aims at discovering manifold structure hidden in high dimensional data. It seeks to discover the intrinsic structure of a given manifold. Notice that when many manifolds are considered, we refer to this as multi-manifold RL [Lee *et al.*, 2016; Torki *et al.*, 2010]. It allows to both preserve the local geometric structure within distinct manifolds while ensuring the discriminability between them [Wu *et al.*, 2020].

When more neural networks transparency is required, the visualisation of latent representations is essential: it allows to develop an intuition about the distance between subsets of data represented by their associated latent manifold representations. Consequently, we consider that this dimension-reduction characteristic is therefore of great practical interest for XAI. Indeed, reducing the complexity due to the high dimensions can strongly contribute in understanding the inner mechanisms of models exploiting the data, but also the role of the data subsets on the models behaviors.

State RL (SRL)

In addition, RL can also concern domains where data are in a low dimensional space. SRL is *"is a particular type of representation learning that aims at building a low-dimensional and meaningful representation of a state space, by processing high-dimensional raw observation data (e.g., learn a position (x, y) from raw image pixels)."* [Heuillet *et al.*, 2021]. This domain is thus particularly suited for learning features in reinforcement learning, robotics and control scenarios. Thus, learning in SRL for an artificial agent is rather related to building a latent model of the environment and the task to perform through interactions [Lesort *et al.*, 2018]. In addition, it has been shown that SRL provides three main advantages for several research domains [Heuillet *et al.*, 2021]:

- The learned features are of low dimensions which improves speed and generalization of deep learning models [Lesort *et al.*, 2017].
- SRL helps improving performance in some reinforcement learning steps such as policy learning [Heuillet *et al.*, 2021].
- Learning representations of states (definition 2.6), actions or policies provide meaning to explain a reinforcement learning algorithms. Indeed, SRL allows to learn representations that capture the variation in the environment generated by the action of the agent [Lesort *et al.*, 2017; Heuillet *et al.*, 2021].

It has been shown that SRL is particularly suitable to make the behavior of an artificial agent and the reasons of this behavior accessible for humans [Lesort *et al.*, 2017; Lesort *et al.*, 2018; Heuillet *et al.*, 2021]. Consequently, we can consider SRL as an example of domains used for explanation goals in reinforcement learning.

4.4 Data-oriented RL approaches

In RL, several approaches tackle the problem of increasing data volumes, their heterogeneity and the multiplicity of their sources. We can consider them as data-oriented approaches and present two of them: the Multi-view RL and the Network RL. We also highlight the link between RL applied to real-world data-oriented problems and XAI domain.

Multi-view RL

In real-world applications, each object can be described by multiple features (definition 2.2) [Xu *et al.*, 2013]. It is thus referred as Multi-view data. These features, also referred to as views, constitute complementary and diverse information of the same data [Xu *et al.*, 2018]. For example, one information can be obtained through multiple sources, which is the case in the application where different people are talking about the same thing. Another example can be an image that is described via a set of visual features such as color, shape and textures. Multi-view RL is thus concerned with the problem of the integration of information from multiple views and uncovers the latent structure shared by multiple views, while preserving the original information and the global meaning [Zhu *et al.*, 2014; Xu *et al.*, 2018]. It has been shown that Multi-view RL can facilitate extracting useful information when developing prediction models [Li *et al.*, 2018] and also helps encoding concepts and semantics in deep neural network [Xu *et al.*, 2018]. Recently, Multi-view RL has been used to design an explainable recommendation system [Gao *et al.*, 2019], where authors claim that *"it is difficult to model the relationships between high-level and low-level features since they have overlapping meaning"*. To overcome this issue, a Multi-view learning approach has been proposed by considering different

levels of features as different views. The learned representation can then be a representation of different levels of features of the input data. Accordingly, we consider that Multi-view RL can be employed for explainability tasks.

Network RL

Network RL is a learning paradigm proposed to analyze networks such as graphs, and thus allows users to deeply understand the hidden features of graphs [Sun *et al.*, 2020]. This domain aims at learning in a low-dimensional space of network vertices (definition 2.8), while preserving the structure of the network topology, the content of the vertices and other information as vertices attributes and links attributes. Network RL can be considered as a dimensionality reduction technique and an intermediate step to solve a target task [Zhang *et al.*, 2020]. Since the information of the original network is preserved in a new vector-based representation, conventional vector-based machine learning algorithms can be applied. Thus, Network analysis and mining tasks become easier as there is no more need to use complex algorithms directly designed for graphs.

Consequently, Network RL has multiple applications such as: vertex classification, link prediction, clustering, visualization and recommendations [Dong *et al.*, 2020; Zhang *et al.*, 2020]. Network RL approaches have been widely applied to information networks [Sun *et al.*, 2020; Zhang *et al.*, 2020] and are becoming increasingly popular for capturing complex relationships in various real-world applications [Yang *et al.*, 2015; Sun *et al.*, 2020; Zhang *et al.*, 2020], such as social networks, citation networks, telecommunication networks, biological networks, recommender systems, *etc.*

In addition, Network RL is essential in the study of heterogeneous information networks (*i.e.* where vertices are of different types), in order to capture semantic proximity between vertices representations [Dong *et al.*, 2020]. Given the high scale of some networks that can range from hundred to billions of vertices and the heterogeneity of information, we believe that Network RL and XAI should be considered together in order to perform efficient and explainable analytical tasks. Also, in related applications, an in depth analysis using XAI techniques and Network RL can help interpreting empirical results and providing a deep understanding of the applied black-box model. To conclude, Network RL should be considered as a dimensionality reduction technique whenever graph-data structure is involved in the design of XAI.

4.5 Discussion: relation between RL and XAI

We have presented several research works in RL (Manifold RL, State RL, Multi-view RL and Network RL) and we next highlight common points between RL and XAI.

First, let us discuss the contribution of the **hierarchical RL** on XAI modeling. Recall that while RL focuses on learning a data representation in order to get a better performance of the AI model [Bengio *et al.*, 2013], XAI is interested in exploring this representation to explain the performance and behavior of the model. This representation varies according to the techniques used in the involved AI models (*e.g.* an artificial agent or a neural network). In the case of deep neural networks models, the hierarchical level of representations is important for XAI, as it allows to extract different types of information that can be used in several ways:

- The study of low-level representations can help to detect important features used by the deep network to make a prediction. This contributes to the explanation and understanding of the deep network by determining features involved in a particular output (*i.e.* a prediction).
- The study of high-level representations can help to detect groups of features involved in a prediction, and how and where a deep neural architecture deals with these groups. This is interesting to explain relevant hidden information and their location within the architecture.

For example, a hierarchical multi-scale deep recurrent network approach has been proposed for data sequences [Chung *et al.*, 2016]: in order to discover temporal dependencies in data, the latent hierarchical structure in the sequences has been exploited without using explicit boundary information. Accordingly, we consider that the hierarchical structure of the latent representations is an important characteristic of deep networks in order to propose a model-specific XAI modeling.

Second, we focus now on the contribution of **problems-oriented** and **data-oriented** RL approaches discussed above on the explainability of AI models.

- Recall that for high-dimensional data sets, **Manifold RL** allows to perform dimension reduction in the latent space while preserving the distance or similarities between data. Consequently, one of the main advantages is that visualisation of the data representation inside the latent space allows to get a better intuition and understanding of the inner mechanisms of models.
- Recall that in reinforcement learning, **SRL** allows to explicit the agent state changes while performing a task in a given environment. This is similar to the XAI objective as it makes the behavior of an artificial agent explicit and more intelligible for a given target audience. Also, recent works have mentioned that State RL can be viewed as a mean for XAI in reinforcement learning [Heuillet *et al.*, 2021]. Other works describe State RL as an approach for robotics and control scenarios that provides easier interpretation of the variation in the environment [Lesort *et al.*, 2017]. Consequently, we can consider that the goals of SRL are in line with those of XAI.
- Through the presentation of **Multi-view RL** and **Network RL** in section 4.4, we have shown that real-world applications of RL techniques that can be more specific to a particular data type or data organisation, are also linked to XAI. Indeed, an AI model can learn from multiple data sets of complex data representation such as networks (*e.g.* social network modeling, biological networks). The complexity of the learned data can also impact the behavior of the AI model. Consequently, this allows us to conclude that adopting RL approaches that take into account the type of learned data, is a way to make AI models more explicit and explainable.

Figure 4 summarizes the above conclusions and questions tackled throughout the section 4. Table 2 summarizes RL domains and some examples of application domains.

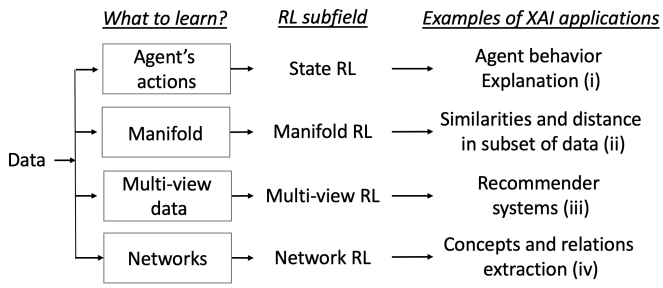

Figure 4: Questions addressed throughout the paper in section 4. Associated reference to each example: (i) [Madumal *et al.*, 2020], (ii) [Torki *et al.*, 2010], (iii) [Gao *et al.*, 2019], (iv) [Qi *et al.*, 2020]

Approach	Non-exhaustive examples of application domain
RL	Speech recognition [Liu *et al.*, 2020] Object recognition [Wang *et al.*, 2020] NLP [Mikolov *et al.*, 2013; Bérard *et al.*, 2016]
State RL	Robotics [Lesort *et al.*, 2017] Numerical artificial agent [Madumal *et al.*, 2020]
Manifold RL	Data mining [Torki *et al.*, 2010]
Multi-view RL	Concept learning [Xu *et al.*, 2018] Image processing [Su *et al.*, 2011] Recommender systems explainability [Gao *et al.*, 2019]
Network RL	Networks of concepts [Yang *et al.*, 2015; Qi *et al.*, 2020] Identification of genes in biology [Ietswaart *et al.*, 2021] Community detection in social networks [Tu *et al.*, 2018]

Table 2: A Summary of RL approaches, examples of application domains (NLP stands for Natural language processing).

5 Discussion and conclusion

We now summarize the highlighted points presented in previous sections. We also present promising directions related to the XAI domain. Since our paper is a multidisciplinary one at the crossroad of several domains, we have **first** (in section 2) centralized and clarified definitions of several concepts, that could indeed seem basic and well-known to involved AI experts, but are important to bridge the discussed domains. A special focus has been made on latent space, latent representation and hierarchical representation which are essential for knowledge extraction in deep neural networks and thus in XAI. To the best of our knowledge, no previous work has established a clear definition of these concepts for XAI community. This is necessary to allow the collaboration between the different domains necessary to build XAI. **Second**, we analysed and highlighted the existence of relations between Knowledge domains (KDP, KR), RL and XAI.

As we have shown in section 1, the goal of XAI is to convey the most semantically complete explanation to a target audience in order to answer a particular question within a given context. This explanation should take into account two important points: (i) the prior knowledge of the target audience regarding the application context, and (ii) the technical aspects of the AI used model that provided solutions to a specific task, and that thus contributed, due to its complexity/opacity, to the emergence of the question behind the need of XAI, *i.e* in short, "What are the reasons behind the results and/or how the AI model reaches these results?".

We consider that XAI is technically at the crossroad of at least two domains: (i) KDP and KR when viewed from a human perspective, and (ii) RL that tackles implicitly the same objectives as XAI, from a technical and algorithmic perspectives. KDP, KR, RL domains, while distinct, are overlapped. They do and should have an explicit impact on XAI approaches:

- First, as we have previously mentioned, several XAI approaches are indirectly inspired by the domain of Knowledge (KDP, KR and data mining) as both tend to express information from data. However, it is important to recall that, in XAI the input data reflects the internal mechanisms of the AI model, its predictions, and/or its behavior. The evolution of the Knowledge domain is therefore an inspiration area for XAI.
- Second, the development of AI approaches and in particular of deep learning, has blurred the boundaries between KR and RL, since several KR approaches involve RL and deep learning. In addition, recall that while RL is interested in features modeling for algorithmic issues (performance, dimensionality, *etc.*), XAI is interested in features since it contributes to explicit the inner mechanisms behind the results. This implies that KR, RL and XAI are indeed interested in the data representation in order to answer different but related questions. We thus consider that, in order to make a significant progress, XAI future works should not forget KR and RL past and recent works as inspirations.

KDP, KR and RL have been extensively confronted with, first, issues related to providing a data-driven explanation to different stakeholders according to their expectations and context, and second, issues related to biases and fairness in AI [Nelson, 2019]. This highlights the human significant role on data processing and bias detection in AI towards XAI. We believe that this review is all the more topical and important as works about the alliance between symbolic AI and connectionist AI should be more and more important in the next years[3], *e.g.* injecting a priori knowledge into neural networks to limit unethical AI [Goebel *et al.*, 2018] and biases [Gordon and Desjardins, 1995; Leavy, 2018; Lepri *et al.*, 2018; Nelson, 2019]. We are convinced that very promising directions can be taken in XAI future works by taking advantage of KDP, KR and RL development to help design ethical, unbiased and human-centered XAI. To conclude, we point out that other domains, not discussed in this paper, also impact XAI directions such as cognitive psychology [Le Saux *et al.*, 2002], cognitive sciences for biases studies [Soleimani *et al.*, 2021], social sciences [Miller, 2019] and Human Machine Interaction field [Le Saux *et al.*, 1999; Mueller *et al.*, 2021; Ehsan *et al.*, 2021].

6 Acknowledgments

Thanks to the *Conseil régional de Bretagne* and the *European Union* for funding this work via the *FEDER* program.

[3]The alliance of symbolic AI and connectionist approaches have been proposed a long time ago, *e.g.* [Honavar, 1995].

References

[Abraham, 2005] Ajith Abraham. Artificial neural networks. *Handbook of measuring system design*, 2005.

[Adadi and Berrada, 2020] Amina Adadi and Mohammed Berrada. Explainable ai for healthcare: From black box to interpretable models. In *Embedded Systems and Artificial Intelligence*, pages 327–337. Springer, 2020.

[Ahn *et al.*, 2016] Sungjin Ahn, Heeyoul Choi, Tanel Pärnamaa, and Yoshua Bengio. A neural knowledge language model. *arXiv preprint arXiv:1608.00318*, 2016.

[Anand and Büchner, 1998] Sarabjot S Anand and Alex G Büchner. *Decision support using data mining*. Financial Times Management, 1998.

[Augusto *et al.*, 2017] J Augusto, Asier Aztiria, Dean Kramer, and Unai Alegre. A survey on the evolution of the notion of context-awareness. *Applied Artificial Intelligence*, 31(7-8):613–642, 2017.

[Barredo Arrieta *et al.*, 2020] Alejandro Barredo Arrieta, Natalia Diaz-Rodriguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador Garcia, Sergio Gil-Lopez, Daniel Molina, Richard Benjamins, Raja Chatila, and Francisco Herrera. Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, 58:82–115, 2020.

[Barto and Sutton, 1995] Andrew G Barto and Richard S Sutton. Reinforcement learning. *Handbook of brain theory and neural networks*, pages 804–809, 1995.

[Bengio *et al.*, 2013] Y. Bengio, A. Courville, and P. Vincent. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1798–1828, 2013.

[Bengio, 2009] Yoshua Bengio. *Learning deep architectures for AI*. Now Publishers Inc, 2009.

[Bérard *et al.*, 2016] Alexandre Bérard, Christophe Servan, Olivier Pietquin, and Laurent Besacier. Multivec: a multilingual and multilevel representation learning toolkit for NLP. In *The 10th edition of the Language Resources and Evaluation Conference*, 2016.

[Bishop, 2006] Christopher M Bishop. *Pattern recognition and machine learning*. springer, 2006.

[Brenon *et al.*, 2018] Alexis Brenon, François Portet, and Michel Vacher. Context feature learning through deep learning for adaptive context-aware decision making in the home. In *2018 14th International Conference on Intelligent Environments (IE)*, pages 32–39. IEEE, 2018.

[Brézillon, 1999] Patrick Brézillon. Context in Artificial Intelligence: I. A survey of the literature. *Computers and artificial intelligence*, 18:321–340, 1999.

[Buchanan and Shortliffe, 1984] Bruce G Buchanan and Edward H Shortliffe. Rule-based expert systems: the MYCIN experiments of the stanford heuristic programming project. 1984.

[Burgt, 2020] Joost van der Burgt. Explainable AI in banking. *Journal of Digital Banking*, 4(4):344–350, 2020.

[Cabena *et al.*, 1998] Peter Cabena, Pablo Hadjinian, Rolf Stadler, Jaap Verhees, and Alessandro Zanasi. *Discovering data mining: from concept to implementation*. Prentice-Hall, Inc., 1998.

[Capone and Bertolaso, 2020] Luca Capone and M. Bertolaso. A philosophical approach for a human-centered explainable AI. In *The Italian Workshop on Explainable Artificial Intelligence, the 19th International Conference of the Italian Association for Artificial Intelligence*, 2020.

[Cayton, 2005] Lawrence Cayton. Algorithms for manifold learning. *Univ. of California at San Diego Tech. Rep*, pages 1–17, 2005.

[Cho *et al.*, 2014] Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using RNN encoder–decoder for statistical machine translation. In *The 18th Empirical Methods in Natural Language Processing*, pages 1724–1734, Doha, Qatar, October 2014. Association for Computational Linguistics.

[Chung *et al.*, 2016] Junyoung Chung, Sungjin Ahn, and Yoshua Bengio. Hierarchical multiscale recurrent neural networks. *arXiv preprint arXiv:1609.01704*, 2016.

[Cios *et al.*, 2007] Krzysztof J Cios, Witold Pedrycz, Roman W Swiniarski, and Lukasz Andrzej Kurgan. *Data mining: a knowledge discovery approach*. Springer Science & Business Media, 2007.

[Das, 2003] Amit Das. Knowledge representation. In Hossein Bidgoli, editor, *Encyclopedia of Information Systems*, pages 33–41. Elsevier, New York, 2003.

[Davis *et al.*, 1993] Randall Davis, Howard Shrobe, and Peter Szolovits. What is a knowledge representation? *AI magazine*, 14(1):17–17, 1993.

[Davis, 2015] Ernest Davis. Knowledge representation. In *International Encyclopedia of the Social & Behavioral Sciences: Second Edition*, pages 98–104. Elsevier Inc., 2015.

[Dietterich *et al.*, 2008] Thomas G Dietterich, Pedro Domingos, Lise Getoor, Stephen Muggleton, and Prasad Tadepalli. Structured machine learning: the next ten years. *Machine Learning*, 73(1):3, 2008.

[Dolgikh, 2018] Serge Dolgikh. Spontaneous concept learning with deep autoencoder. *International Journal of Computational Intelligence Systems*, 12(1):1, 2018.

[Dong *et al.*, 2020] Yuxiao Dong, Ziniu Hu, Kuansan Wang, Yizhou Sun, and Jie Tang. Heterogeneous network representation learning. In *The 29th International Joint Conference on Artificial Intelligence*, pages 4861–4867, 2020.

[Ehsan *et al.*, 2021] Upol Ehsan, Philipp Wintersberger, Q Vera Liao, Martina Mara, Marc Streit, Sandra Wachter, Andreas Riener, and Mark O Riedl. Operationalizing

human-centered perspectives in explainable AI. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–6, 2021.

[Fayyad *et al.*, 1996] Usama Fayyad, Gregory Piatetsky-Shapiro, and Padhraic Smyth. From data mining to knowledge discovery in databases. *AI magazine*, 17(3):37–37, 1996.

[Fazi, 2020] M Beatrice Fazi. Beyond human: Deep learning, explainability and representation. *Theory, Culture & Society*, pages 1–23, 2020.

[Fletcher *et al.*, 1991] Peter Fletcher, Hughes Hoyle, and C Wayne Patty. *Foundations of discrete mathematics*. Brooks/Cole, 1991.

[Gao *et al.*, 2019] Jingyue Gao, Xiting Wang, Yasha Wang, and Xing Xie. Explainable recommendation through attentive multi-view learning. *The 33th AAAI Conference on Artificial Intelligence*, 33(01):3622–3629, Jul. 2019.

[Gilpin *et al.*, 2018] Leilani H Gilpin, David Bau, Ben Z Yuan, Ayesha Bajwa, Michael Specter, and Lalana Kagal. Explaining explanations: An overview of interpretability of machine learning. In *The 5th IEEE International Conference on data science and advanced analytics*, pages 80–89. IEEE, 2018.

[Goebel *et al.*, 2018] Randy Goebel, Ajay Chander, Katharina Holzinger, Freddy Lecue, Zeynep Akata, Simone Stumpf, Peter Kieseberg, and Andreas Holzinger. Explainable ai: the new 42? In *International cross-domain conference for machine learning and knowledge extraction*, pages 295–303. Springer, 2018.

[Gordon and Desjardins, 1995] Diana F Gordon and Marie Desjardins. Evaluation and selection of biases in machine learning. *Machine learning*, 20(1-2):5–22, 1995.

[Grazzini and Pantisano, 2015] Jacopo Grazzini and Francesco Pantisano. Collaborative research-grade software for crowd-sourced data exploration: from context to practice - part I: Guidelines for scientific evidence provision for policy support based on big data and open technologies. EUR 27094. *Luxembourg: Publications Office of the European Union*, (JRC94504), 2015.

[Guarino *et al.*, 2009] Nicola Guarino, Daniel Oberle, and Steffen Staab. What is an ontology? In *Handbook on ontologies*, pages 1–17. Springer, 2009.

[Guidotti *et al.*, 2018] Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi. A survey of methods for explaining black box models. *ACM computing surveys (CSUR)*, 51(5):93, 2018.

[Gunning, 2017] David Gunning. Explainable artificial intelligence (XAI). *Defense Advanced Research Projects Agency (DARPA)*, 2(2), 2017.

[Hagedorn *et al.*, 2020] Thomas Hagedorn, Mary Bone, Benjamin Kruse, Ian Grosse, and Mark Blackburn. Knowledge representation with ontologies and semantic web technologies to promote augmented and artificial intelligence in systems engineering. *Insight*, 23(1):15–20, 2020.

[Hamilton, 2020] William L Hamilton. Graph representation learning. *Synthesis Lectures on Artifical Intelligence and Machine Learning*, 14(3):1–159, 2020.

[He *et al.*, 2021] Xin He, Kaiyong Zhao, and Xiaowen Chu. Automl: A survey of the state-of-the-art. *Knowledge-Based Systems*, 212:106622, 2021.

[Heuillet *et al.*, 2021] Alexandre Heuillet, Fabien Couthouis, and Natalia Diaz-Rodriguez. Explainability in deep reinforcement learning. *Knowledge-Based Systems*, 214:106685, 2021.

[Hochreiter and Schmidhuber, 1997] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[Hollister *et al.*, 2017] Debra L Hollister, Avelino Gonzalez, and James Hollister. Contextual reasoning in human cognition and the implications for artificial intelligence systems. In *The International and Interdisciplinary Conference on Modeling and Using Context*, pages 599–608. Springer, 2017.

[Honavar, 1995] Vasant Honavar. Symbolic artificial intelligence and numeric artificial neural networks: towards a resolution of the dichotomy. In *Computational architectures integrating neural and symbolic processes*, pages 351–388. Springer, 1995.

[Ietswaart *et al.*, 2021] Robert Ietswaart, Benjamin M Gyori, John A Bachman, Peter K Sorger, and L Stirling Churchman. Genewalk identifies relevant gene functions for a biological context using network representation learning. *Genome biology*, 22(1):1–35, 2021.

[Kingma and Welling, 2014] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. In *The 2nd International Conference on Learning Representations*, 2014.

[Le Saux *et al.*, 1999] Elisabeth Le Saux, Philippe Lenca, Philippe Picouet, and Jean-Pierre Barthélemy. An anthropocentric tool for decision making support. In *The International Joint Conference on Artificial Intelligence*, volume 16, pages 338–343. Lawrence Erlbaum Associates LTD, 1999.

[Le Saux *et al.*, 2002] Elisabeth Le Saux, Philippe Lenca, and Philippe Picouet. Dynamic adaptation of rules bases under cognitive constraints. *European Journal of Operational Research*, 136(2):299 – 309, January 2002.

[Leavy, 2018] Susan Leavy. Gender bias in artificial intelligence: The need for diversity and gender theory in machine learning. In *Proceedings of the 1st international workshop on gender equality in software engineering*, pages 14–16, 2018.

[LeCun *et al.*, 2015] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553):436–444, 2015.

[Lee *et al.*, 2016] Chan-Su Lee, Ahmed Elgammal, and Marwan Torki. Learning representations from multiple manifolds. *Pattern Recognition*, 50:74–87, 2016.

[Lenca, 2002] Philippe Lenca. Human centered processes. *European Journal of Operational Research*, 136(2):231–232, 2002.

[Lepri *et al.*, 2018] Bruno Lepri, Nuria Oliver, Emmanuel Letouzé, Alex Pentland, and Patrick Vinck. Fair, transparent, and accountable algorithmic decision-making processes. *Philosophy & Technology*, 31(4):611–627, 2018.

[Lesort *et al.*, 2017] Timothée Lesort, Mathieu Seurin, Xinrui Li, Natalia Díaz-Rodríguez, and David Filliat. Unsupervised state representation learning with robotic priors: a robustness benchmark. *arXiv preprint arXiv:1709.05185*, 2017.

[Lesort *et al.*, 2018] Timothée Lesort, Natalia Díaz-Rodríguez, Jean-François Goudou, and David Filliat. State Representation Learning for Control: An Overview. *Neural Networks*, 108:379–392, December 2018.

[Li *et al.*, 2018] Yingming Li, Ming Yang, and Zhongfei Zhang. A survey of multi-view representation learning. *IEEE Transactions on Knowledge and Data Engineering*, 31(10):1863–1883, 2018.

[Lim *et al.*, 2009] Brian Y Lim, Anind K Dey, and Daniel Avrahami. Why and why not explanations improve the intelligibility of context-aware intelligent systems. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 2119–2128, 2009.

[Lin *et al.*, 2018] Yankai Lin, Xu Han, Ruobing Xie, Zhiyuan Liu, and Maosong Sun. Knowledge representation learning: A quantitative review. *CoRR*, abs/1812.10901, 2018.

[Liu *et al.*, 2020] Alexander H Liu, Tao Tu, Hung-yi Lee, and Lin-shan Lee. Towards unsupervised speech recognition and synthesis with quantized speech representation learning. In *The 45th IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 7259–7263. IEEE, 2020.

[Lundberg and Lee, 2017] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems 30*, pages 4765–4774. Curran Associates, Inc., 2017.

[Madumal *et al.*, 2020] Prashan Madumal, Tim Miller, Liz Sonenberg, and Frank Vetere. Explainable reinforcement learning through a causal lens. In *The 34th AAAI Conference on Artificial Intelligence*, pages 2493–2500, Apr. 2020.

[Malhotra and Nair, 2015] Meenakshi Malhotra and TR Gopalakrishnan Nair. Evolution of knowledge representation and retrieval techniques. *International Journal of Intelligent Systems and Applications*, 7(7):18–28, 2015.

[McCulloch and Pitts, 1943] Warren S McCulloch and Walter Pitts. A logical calculus of the ideas immanent in nervous activity. *The bulletin of mathematical biophysics*, 5(4):115–133, 1943.

[Mikolov *et al.*, 2013] Tomás Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. In Yoshua Bengio and Yann LeCun, editors, *1st International Conference on Learning Representations, Scottsdale, Arizona, USA, May 2-4, 2013, Workshop Track Proceedings*, 2013.

[Miller, 2019] Tim Miller. Explanation in artificial intelligence: Insights from the social sciences. *Artificial intelligence*, 267:1–38, 2019.

[Mueller *et al.*, 2021] Shane T Mueller, Elizabeth S Veinott, Robert R Hoffman, Gary Klein, Lamia Alam, Tauseef Mamun, and William J Clancey. Principles of explanation in human-ai systems. *arXiv preprint arXiv:2102.04972*, 2021.

[Nelson, 2019] Gregory S Nelson. Bias in artificial intelligence. *North Carolina medical journal*, 80(4):220–222, 2019.

[Nguyen *et al.*, 2021] Thao Nguyen, Maithra Raghu, and Simon Kornblith. Do wide and deep networks learn the same things? Uncovering how neural network representations vary with width and depth. In *9th International Conference on Learning Representations, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021.

[Puiutta and Veith, 2020] Erika Puiutta and Eric M. S. P. Veith. Explainable reinforcement learning: A survey. In Andreas Holzinger, Peter Kieseberg, A Min Tjoa, and Edgar R. Weippl, editors, *Machine Learning and Knowledge Extraction - 4th IFIP TC 5, TC 12, WG 8.4, WG 8.9, WG 12.9 International Cross-Domain Conference, CD-MAKE 2020, Dublin, Ireland, August 25-28, 2020, Proceedings*, volume 12279 of *Lecture Notes in Computer Science*, pages 77–95. Springer, 2020.

[Qi *et al.*, 2020] Zhaobo Qi, Shuhui Wang, Chi Su, Li Su, Qingming Huang, and Qi Tian. Towards more explainability: Concept knowledge mining network for event recognition. In *The 28th ACM International Conference on Multimedia*, page 3857–3865, New York, NY, USA, 2020. Association for Computing Machinery.

[Ribeiro *et al.*, 2016] Marco Túlio Ribeiro, Sameer Singh, and Carlos Guestrin. "Why Should I Trust you?": Explaining the predictions of any classifier. In *The 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144, 2016.

[Roberts *et al.*, 2018a] Adam Roberts, Jesse Engel, Colin Raffel, Curtis Hawthorne, and Douglas Eck. A hierarchical latent vector model for learning long-term structure in music. In *The 35th International Conference on Machine Learning*, pages 4364–4373. PMLR, 2018.

[Roberts *et al.*, 2018b] Adam Roberts, Jesse H Engel, Sageev Oore, and Douglas Eck. Learning latent representations of music to generate interactive musical palettes. In *IUI Workshops*, 2018.

[Rojat *et al.*, 2021] Thomas Rojat, Raphaël Puget, David Filliat, Javier Del Ser, Rodolphe Gelin, and Natalia Díaz-Rodríguez. Explainable artificial intelligence (XAI) on Time Series Data: A Survey. *arXiv preprint arXiv:2104.00950*, 2021.

[Rumelhart *et al.*, 1986] David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. Learning representations by back-propagating errors. *Nature*, 323(6088):533–536, 1986.

[Sallinger *et al.*, 2020] Emanuel Sallinger, Sahar Vahdati, and Mojtaba Nayyeri. Workshop knowledge representation & representation learning, 2020.

[Schlegel *et al.*, 2019] Udo Schlegel, Hiba Arnout, Mennatallah El-Assady, D. Oelke, and D. Keim. Towards a rigorous evaluation of XAI methods on time series. *The 2019 IEEE/CVF International Conference on Computer Vision Workshop*, pages 4197–4201, 2019.

[Shapiro, 2006] Stuart C Shapiro. Knowledge representation. *Encyclopedia of cognitive science*, 2006.

[Shortliffe, 1974] Edward H Shortliffe. A rule-based computer program for advising physicians regarding antimicrobial therapy selection. In *Proceedings of the 1974 annual ACM conference-Volume 2*, pages 739–739, 1974.

[Soleimani *et al.*, 2021] Melika Soleimani, Ali Intezari, Nazim Taskin, and David Pauleen. Cognitive biases in developing biased Artificial Intelligence recruitment system. In *Proceedings of the 54th Hawaii International Conference on System Sciences*, page 5091, 2021.

[Su *et al.*, 2011] Ya Su, Yun Fu, Xinbo Gao, and Qi Tian. Discriminant learning through multiple principal angles for visual recognition. *IEEE transactions on image processing*, 21(3):1381–1390, 2011.

[Sun *et al.*, 2020] Ke Sun, Lei Wang, Bo Xu, Wenhong Zhao, Shyh Wei Teng, and Feng Xia. Network representation learning: From traditional feature learning to deep learning. *IEEE Access*, 8:205600–205617, 2020.

[Torki *et al.*, 2010] Marwan Torki, Ahmed Elgammal, and Chan Su Lee. Learning a joint manifold representation from multiple data sets. In *The 20th International Conference on Pattern Recognition*, pages 1068–1071. IEEE, 2010.

[Tu *et al.*, 2018] Cunchao Tu, Xiangkai Zeng, Hao Wang, Zhengyan Zhang, Zhiyuan Liu, Maosong Sun, Bo Zhang, and Leyu Lin. A unified framework for community detection and network representation learning. *IEEE Transactions on Knowledge and Data Engineering*, 31(6):1051–1065, 2018.

[Van Fraassen, 1988] Bas Van Fraassen. The pragmatic theory of explanation. *Theories of Explanation*, 8:135–155, 1988.

[Vilone and Longo, 2020] Giulia Vilone and Luca Longo. Explainable artificial intelligence: a systematic review. *arXiv preprint arXiv:2006.00093*, 2020.

[Walton, 2004] Douglas Walton. A new dialectical theory of explanation. *Philosophical Explorations*, 7(1):71–89, 2004.

[Wang *et al.*, 2020] Jingdong Wang, Ke Sun, Tianheng Cheng, Borui Jiang, Chaorui Deng, Yang Zhao, Dong Liu, Yadong Mu, Mingkui Tan, Xinggang Wang, et al. Deep high-resolution representation learning for visual recognition. *IEEE transactions on pattern analysis and machine intelligence (accepted)*, 2020.

[Wirth, 2000] Rüdiger Wirth. Crisp-dm: Towards a standard process model for data mining. In *The 4th International Conference on the Practical Application of Knowledge Discovery and Data Mining*, pages 29–39, 2000.

[Wu *et al.*, 2020] Lirong Wu, Zicheng Liu, Zelin Zang, Jun Xia, Siyuan Li, and Stan Z Li. Deep clustering and representation learning that preserves geometric structures. *arXiv e-prints*, pages arXiv–2009, 2020.

[Xie *et al.*, 2018] Ruobing Xie, Zhiyuan Liu, Fen Lin, and Leyu Lin. Does William Shakespeare REALLY Write Hamlet? Knowledge Representation Learning With Confidence. In Sheila A. McIlraith and Kilian Q. Weinberger, editors, *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence*, pages 4954–4961. AAAI Press, 2018.

[Xu *et al.*, 2013] Chang Xu, Dacheng Tao, and Chao Xu. A survey on multi-view learning. *arXiv preprint arXiv:1304.5634*, 2013.

[Xu *et al.*, 2018] Cai Xu, Ziyu Guan, Wei Zhao, Yunfei Niu, Quan Wang, and Zhiheng Wang. Deep multi-view concept learning. In *The 27th International Joint Conference on Artificial Intelligence*, pages 2898–2904, 2018.

[Yang *et al.*, 2015] Cheng Yang, Zhiyuan Liu, Deli Zhao, Maosong Sun, and Edward Y Chang. Network representation learning with rich text information. In *The 24th International Joint Conference on Artificial Intelligence*, volume 2015, pages 2111–2117, 2015.

[Yin *et al.*, 2016] Jun Yin, Xin Jiang, Zhengdong Lu, Lifeng Shang, Hang Li, and Xiaoming Li. Neural generative question answering. In *The 25th International Joint Conference on Artificial Intelligence*, 2016.

[Zhang *et al.*, 2011] Zhenyue Zhang, Jing Wang, and Hongyuan Zha. Adaptive manifold learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(2):253–265, 2011.

[Zhang *et al.*, 2016] Fuzheng Zhang, Nicholas Jing Yuan, Defu Lian, Xing Xie, and Wei-Ying Ma. Collaborative knowledge base embedding for recommender systems. In *The 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 353–362, 2016.

[Zhang *et al.*, 2020] D. Zhang, J. Yin, X. Zhu, and C. Zhang. Network representation learning: A survey. *IEEE Transactions on Big Data*, 6(01):3–28, jan 2020.

[Zhong *et al.*, 2016] Guoqiang Zhong, Li-Na Wang, Xiao Ling, and Junyu Dong. An overview on data representation learning: From traditional feature learning to recent deep learning. *The Journal of Finance and Data Science*, 2(4):265–278, 2016.

[Zhu *et al.*, 2014] Zhenfeng Zhu, Linlin Du, Lei Zhang, and Yao Zhao. Shared subspace learning for latent representation of multi-view data. *Journal of Information Hiding and Multimedia Signal Processing*, 5(3):546–554, 2014.