Data Intelligence Just Accepted MS.
https://doi.org/10.1162/dint_a_00119

Provenance documentation to enable explainable and trustworthy AI: A literature review

Amruta Kale[1], Tin Nguyen[2], Frederick C. Harris Jr.[2], Chenhao Li[1], Jiyin Zhang[1], Xiaogang Ma[1,*]

[1] Department of Computer Science, University of Idaho, Moscow, ID 83844, USA

[2] Department of Computer Science and Engineering, University of Nevada, Reno, Reno, NV 89557, USA

* Correspondence author. Email: max@uidaho.edu; Tel.: +1 208 885 1547

Abstract:

Recently artificial intelligence (AI) and machine learning (ML) models have demonstrated remarkable progress with applications developed in various domains. It is also increasingly discussed that AI and ML models and applications should be transparent, explainable, and trustworthy. Accordingly, the field of Explainable AI (XAI) is expanding rapidly. XAI holds substantial promise for improving trust and transparency in AI-based systems by explaining how complex models such as the deep neural network (DNN) produces their outcomes. Moreover, many researchers and practitioners consider that using provenance to explain these complex models will help improve transparency in AI-based systems. In this paper, we conduct a systematic literature review of provenance, XAI, and trustworthy AI (TAI) to explain the fundamental concepts and illustrate the potential of using provenance as a medium to help accomplish explainability in AI-based systems. Moreover, we also discuss the patterns of recent developments

in this area and offer a vision for research in the near future. We hope this literature review will serve as a starting point for scholars and practitioners interested in learning about essential components of provenance, XAI, and TAI.

Keywords: Explainable AI, Trustworthy AI, Provenance documentation, Workflow platforms, Data science

1 Introduction

Over the past decade, the rapid rise of applications in artificial intelligence (AI) has raised the discussion of explainable AI (XAI) and trustworthy AI (TAI) among data science practitioners (Wing, 2020). We have seen remarkable progress in AI algorithms and facilities for high-performance computation, and applications of AI are thriving in various domains, such as virtual assistants, healthcare, autonomous vehicles, criminal justice, human resource, and environmental science. In many applications, the results generated by AI/ML models have a huge impact on human decision-making. However, existing models are insufficient to certify how and why the results were obtained, which leads to growing concerns that these AI/ML models are unfair, opaque, or non-intuitive (Goodman and Flaxman, 2017). For example, ML and deep learning (DL) are the most representative technologies in AI and are widely used by data science practitioners. ML is a powerful tool and can identify patterns and examine correlations on large datasets. DL is a subset of ML that achieves great power and flexibility (Goodfellow et al., 2016). It uses a vast amount of labeled data and multiple layers of algorithms to imitate the neural network in our brain, with the aim to achieve human-like cognitive abilities. The most representative DL technology is the artificial neural network (ANN) or deep neural network (DNN). DNN comprises a large number of neurons or nodes with each layer. These nodes are interconnected in a complex

manner and activate multiple combinations at each layer. However, it is debatable how this complex network works and derives its output which leads to the "black-box" problem (Castelvecchi, 2016). Although these models perform complex computational tasks with high predictive accuracy, we need to ensure that the steps, workflows, and results of these models are transparent, interpretable, unbiased, and trustworthy. One of the approaches for increasing transparency is to explain these complex models through XAI, which in turn is a feasible step in building TAI (Adadi and Berrada, 2018).

The goal of XAI is to provide algorithmic transparency that can be understood by the average human being (Ribeiro et al., 2016). XAI will help to answer questions like how the system made certain predictions, why the system fails, or what biases are present in the system or data (Guidotti et al., 2018; Murdoch et al., 2019). However, not all AI applications need explanation. Some practitioners and academics discussed that explaining a black-box model is difficult to achieve or perhaps unnecessary. Instead, they suggested that these models should be designed inherently interpretable (Rudin, 2018, 2019; Rudin and Radin, 2019). This approach is highly debatable, as in most of the applications the accurate predictive solutions are provided by complex ML models. Some of the ML models such as rule-based learning, K-nearest neighbor, and linear regression have high interpretability and their workflows are easy to understand. However, many other AI models such as DNN, support vector machine (SVM), and Bayesian models have complex structures and workflows, which are mysterious to the outside observers. On some occasions, even the programmers of these models are incapable of explaining why a model behaves in a certain way and generates a specific output. With the growing use of AI applications in every aspect of our modern life, there is also an increased risk of unanticipated behavior. The danger is in creating

and using decisions that are not justifiable, legitimate, or that merely do not allow obtaining detailed explanations of their behavior. In that sense, XAI and TAI will be qualified to reveal the strengths, limitations, and/or weaknesses of AI/ML models. They are also an important means to establish user engagement and trust in AI applications.

The technical approaches for XAI and TAI are under quick development, at which some researchers highlighted that provenance is an evolving field to explain AI-based systems (Liu et al., 2017; Jentzsch and Hochgeschwender, 2019; Frost, 2019). Provenance answers the question of who-what-when-where by documenting the process at each step, such as entities, agents, and activities. By portraying transparency, the documented provenance helps trace back the origin of data, demonstrate the steps of data processing, and determine the trustworthiness of results (Jaigirdar et al., 2019; Amalina et al., 2019; Jaigirdar et al., 2020). Given the non-intuitive nature of many AI/ML algorithms, tracking provenance in AI/ML workflows will be helpful since it is an effective technique to highlight significant components in the process and allows scientists to understand how the result was obtained (Samuel et al., 2020). To achieve repeatability and comparability in AI/ML experiments, one must first understand the metadata and most importantly the provenance of the artifacts in the ML process (Kumar et al., 2016). Very recently, Werder and Balasubramaniam (2021) also suggested that data provenance assists and improve fairness, accountability, transparency, and explainability (FATE) in AI/ML algorithms and enables trust. Several other researchers (Liu et al., 2017; Jentzsch and Hochgeschwender, 2019; Frost, 2019) suggested that provenance documentation is an emerging approach toward XAI and TAI. Nevertheless, the work in this field is still limited and there is no systematic discussion or road map for those topics in multi-disciplinary data science.

We anticipate that provenance documentation is an important factor in building XAI and TAI as it not only provides metadata of a workflow but also confirms the authenticity and reproducibility of results. This paper aims to conduct a literature review of existing research on XAI, TAI, and provenance, with a focus on their applications in data science. We started our literature search by scrutinizing academic papers from Scopus as it is one of the largest and most reliable literature databases for scientific research. The search was conducted based on keywords to select papers. We used generic search strings to get more search results like "explainable ai", "trustworthy ai", "artificial intelligence", "explainable artificial intelligence", "machine learning", and "provenance". Our objective was to focus on recent advances. Therefore, we restricted our search from 2010 to 2020. We followed the standard systematic literature review method with backward and forward snowballing strategies (Wohlin, 2014). Snowballing strategy uses a reference list of the paper or citations of the paper to identify additional papers. The gathered papers were then scanned based on the title, abstract, and keywords to verify whether the reported work includes work on XAI, TAI, and provenance. We did not aim to survey all research papers. Instead, we divided our search based on two standards; 1) selection based on a higher level of citation and 2) high-quality papers including good coverage and technicality in the field. Irrelevant articles were excluded, and the remaining articles were examined in detail to understand whether they provide enough information about the proposed methodology, technical approaches, and results. In addition to the literature found on Scopus, in the review and discussion, we also incorporated a number of other publications that deliver a good definition of fundamental concepts and illustrate successful applications.

The structure of this article is organized as follows. Section 2 introduces the concepts of TAI and XAI. Section 3 uses bibliometric analysis to illustrate the latest work in the fields of provenance, XAI, and TAI and demonstrates the interconnections between them. Section 4 explains how provenance documentation plays a fundamental role in TAI and XAI by analyzing their relationships on a more detailed level. Section 5 discusses a few potential research directions of provenance, TAI, and XAI in the next decade. Finally, Section 6 concludes the paper.

2 Fundamental concepts of XAI and TAI

2.1 Background of explainability and trustworthiness in AI

AI/ML models have achieved rapid progress and worldwide adoption, and many of them can be seen on our streets and at our homes. However, despite the successful AI applications, we still lack a scientific understanding of their workflows. To gain more benefit out of these AI-based systems they first need to explain to humans why they made a certain decision and which important features they considered in the process (Montavon et al., 2017; Adadi and Berrada, 2018; Miller, 2019). There are numerous reasons why these systems should be understandable, interpretable, and explainable. It will not only gain trust in humans but will also give confidence that the system works well. In recent years there have been several controversies where the outcomes generated by AI/ML models were biased or discriminatory (Osoba and Welser IV, 2017; Chen et al., 2019). These models have become so dominant that they are raising doubts about future humanity and demand an explanation. For example, in 2016 Microsoft launched a Twitter bot called “Tay”, which was designed to entertain and engage people. In less than 24 hours, Tay’s talk extended to racist and offensive comments, forcing Microsoft to take it offline (Tennery and Cherelus, 2016; Vincent, 2016). There were even life-threatening incidents caused by AI. In 2015, a self-driving

Tesla was involved in a deadly accident in China when it was in autopilot mode and failed to identify a road-sweeping truck (Boudette, 2016). In another incident reported in 2018, a self-driving Uber killed a woman in Arizona. It turned out that the automatic car's software had no capability to classify an object as a pedestrian until that object was near a crosswalk (Mcfarland, 2018; McCausland, 2019). The IBM Watson system once failed to recommend correct treatments for cancer patients (Ross and Swetlitz, 2018). Also, Amazon's AI recruiting tool displayed a gender bias. It was demonstrated that the new recruiting tool was trained to screen applicants by looking for patterns in applications submitted to the company. The majority of the submissions were from men candidates, reflecting male dominance in the tech industry. Accordingly, the AI recruiting tool trained itself that male candidates were preferable, which eventually led to the gender inequality in its recommendations (Dastin, 2018). There are several more examples mentioned in the literature where AI-based systems malfunctioned (e.g., Tan et al., 2017; Adadi and Berrada, 2018). Accordingly, there is a growing need for tools to check vulnerabilities and flaws in AI-based systems, as well as to help developers and users understand why the machine makes a certain decision.

The basic principle of TAI is to build AI-based systems that are lawful, ethical, and robust to ensure that humans can rely on them (Floridi, 2019; Thiebes et al., 2020; Jain et al., 2020). The key to establish TAI is by using XAI, which refers to the series of frameworks and techniques used to ensure that the results generated by AI-based systems are easily understandable and interpretable to humans (Gunning and Aha, 2019). Explainability plays a crucial role in achieving trust and transparency in AI algorithms. To improve explainability, data science practitioners have developed many approaches and strategic plans on XAI. For example, the National Academies of

Sciences and the Royal Society organized a forum in 2017, which reported that trust, transparency, interpretability, and fairness are the most significant societal challenges in AI-based systems (NAS, 2018). Simultaneously, the Defense Advanced Research Projects Agency (DARPA) funded the "Explainable AI (XAI) Program" to improve the explainability of AI results (Gunning, 2019). Also, in July 2017, "The New Generation Artificial Intelligence Development Plan" was sanctioned by China's State Council, to encourage explainability and extensibility (Roberts et al., 2021). In May 2018, the European Parliament set the law of General Data Protection Regulation (GDPR) to award citizens a "Right to Explanation" in cases where their activities are affected by AI (Goddard, 2017). Soon after that, in June 2018 a High-Level Expert Group on AI (HLEG) was set up in the European Commission to design the guideline for TAI (AI HLEG, 2019). The government of Finland published a final report on Finland's artificial intelligence programs in June 2019 in order to position Finland as a leader in the application of AI (MEAEF, 2019). To encourage public trust and promoting the use of AI in the federal government, the White House signed an executive order on TAI in December 2020 (White House, 2020). Along with those efforts, the topics of XAI and TAI have received great attention in the academic, industrial, and governmental sectors.

Very recently, Wing (2021) outlined research agendas that combine the concepts of trustworthy computing, AI, and formal methods for ensuring trustworthiness. In her view, the previous discussion on trustworthy computing covers a set of topics: reliability, safety, security, privacy, availability, and usability. The AI/ML systems especially DL models add a dimension of complexity to traditional computing systems and raise more topics of interest, such as accuracy, robustness, fairness, accountability, transparency, interpretability/explainability, ethics, and more. She also pointed out that although the ML community takes accuracy as a gold standard, XAI and

TAI will require trade-offs among the topics mentioned above. In recent years, XAI and TAI topics have also been increasingly discussed in workshops and conferences. For instance, the Fairness, Accountability, and Transparency in Machine Learning (FAT/ ML) conference series are a unique venue for those topics (Rakova et al., 2020). The records of search queries and publications also reflect the increasing attention to XAI and TAI. The graph in Figure 1 shows the popularity of keywords on Google Trends from 01/2017 to 12/2020. For the same period, we found 772 publications on Scopus whose title, abstract, or keywords refer to XAI or TAI. Figure 2 shows the distribution of those publications in each year.

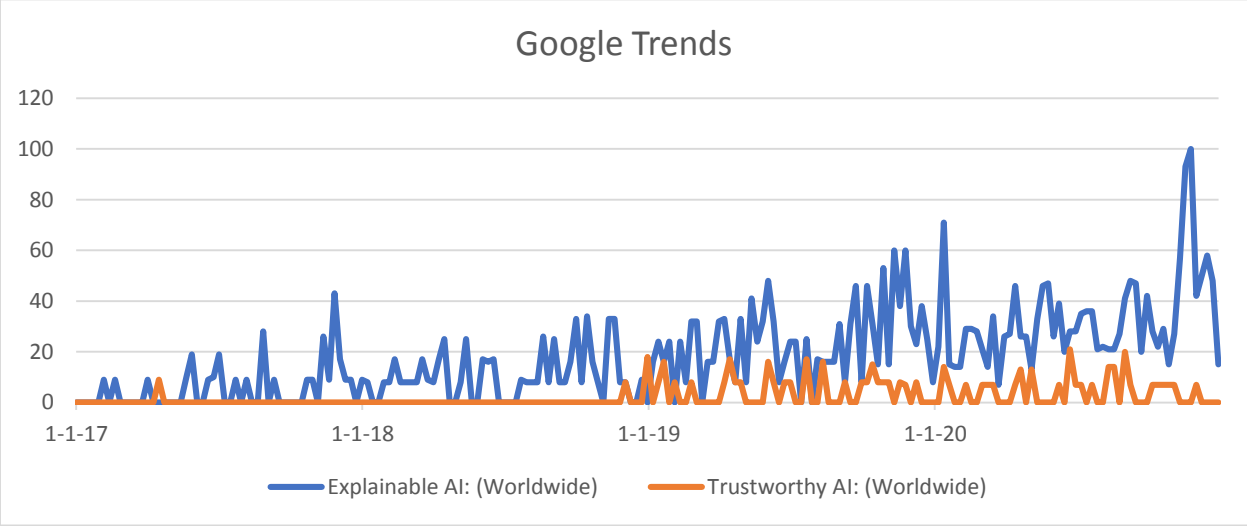

Figure 1: Interest over time (01/01/2017 – 12/31/2020) for the terms "Explainable AI" and "Trustworthy AI" in search queries as shown in Google Trends. The value on the vertical axis is a normalized measure of a topic's popularity among all searches on all topics.

Data Intelligence Just Accepted MS.
https://doi.org/10.1162/dint_a_00119

Document Type	Year of Publication 2017	2018	2019	2020	Grand Total
Article			21	53	74
Article in Press				6	6
Book Chapter				5	5
Conference Paper			57	158	215
Conference Review			1	11	12
Editorial			1		1
Final	5	80	127	234	446
Letter				1	1
Note			1	1	2
Review			3	7	10
Grand Total	5	80	211	476	772

Number of records: 1 – 234

Figure 2: Distribution of publications (01/01/2017 – 12/31/2020) whose title, abstract, or keywords include "Explainable AI" or "Trustworthy AI". This query was used to extract the results from Scopus: (TITLE-ABS-KEY ("Explainable AI") OR TITLE-ABS-KEY ("Trustworthy AI")) AND PUBYEAR > 2016 AND PUBYEAR < 2021. The query was conducted on Aug 1st, 2021.

2.2 Technical approaches for XAI and TAI

There have been several advances in explanation methods and strategies to make AI-based systems more ethical, transparent, and explainable (Singh et al., 2018). In particular, there have been many discussions on technical approaches to enable XAI and TAI in ML models. ML models are classified into two types: transparent and opaque (Belle and Papantonis, 2020). The transparent ML models are recognized as understandable and capable of explaining to some degree by themselves, such as logistics/linear regression, decision tree, k-nearest neighbors, and Bayesian models (Holzinger et al., 2017; Murdoch et al., 2019). These models can fit well when the primary dataset is not complex. In contrast, opaque ML models are "black-box" in nature, making them complex and tricky to understand. Despite obtaining high predictive accuracy, they lack explainability or interpretability of how the results are generated (Montavon et al., 2017; Adadi

and Berrada, 2018). Convolutional neural network (CNN), recurrent neural network (RNN), support vector machine (SVM) and random forest (RF) are the algorithms that fall under opaque models. For instance, RF was initially introduced as a technique to improve accuracy using a single decision tree. In that situation, RF can be treated as a 'transparent' model. However, this technique often suffers from overfitting and poor generalization. To address this issue RF combines multiple trees in which each individual tree is trained on a different part of the training dataset and captures different characteristics to calculate the final outcome. This whole process is far more challenging to explain and lacks interpretability than a single tree, forcing the user to apply a post-hoc explainability approaches to gain more insights from it (Belle and Papantonis, 2020, Arrieta et al., 2020). A post-hoc explainability approach is often employed to extract information about what the model has learned (Guidotti et al., 2018). It means that, when an ML model is unable to explain the intricate method, a separate model is applied to provide an explanation. The post-hoc explainability is categorized into two different techniques: model-agnostic and model-specific (Miller, 2019). The model-agnostic technique can be applied to any type of ML model no matter how complex they are. For instance, some model-agnostic techniques such as Local Interpretable Model Agnostic Explanations (LIME) (Ribeiro et al., 2016) and SHapely Additive exPLanations (SHAP) (Lundberg and Leen, 2017) are widely used to explain DL models. While model-specific technique is only applicable to a single model or a class of models, Tree SHAP (TSHAP) (Lundberg et al., 2020) and Integrated Gradients (IG) (Sundararajan et al., 2017) are some of the popular techniques used for explaining the ML models. When compared with the model-specific techniques, the model-agnostic techniques are more flexible (Ribeiro et al., 2016). Figure 3 depicts the classification of ML models and the corresponding XAI approaches, in which

we have taken the motivation from Arrieta et al. (2020) and Belle and Papantonis (2020), but we adapted the organizational structure to better match the topics discussed here.

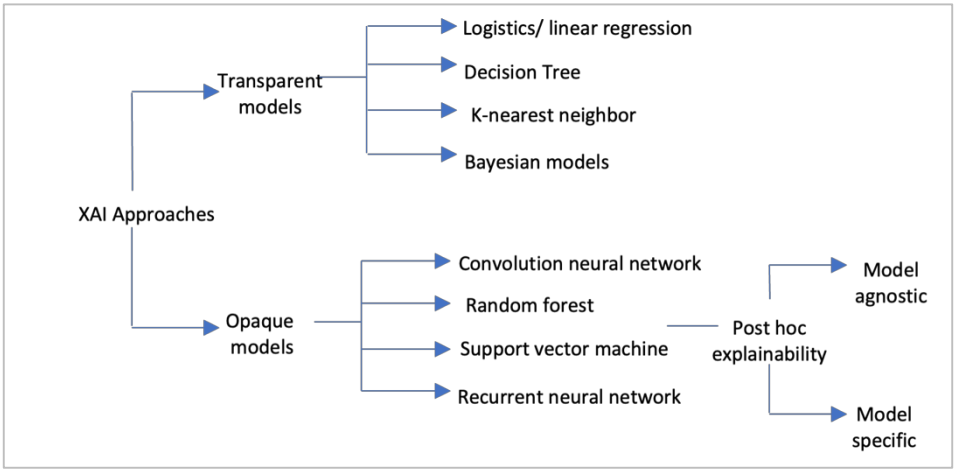

Figure 3: Classification of ML models and XAI approaches

Although these XAI approaches can generate results to explain an ML model, many metadata and context information are still missing. To increase transparency and explainability in AI-based systems, applying provenance documentation can be a complementary technology to the existing XAI approaches (cf. Singh et al., 2018; Jentzsch and Hochgeschwender, 2019). Provenance documentation shows promise in increasing transparency as it can be used for many purposes, such as understanding how data were collected, determining ownership and rights, tracing steps in data analysis, and making judgments about resources to use. Section 3 presents a detailed bibliometric analysis to demonstrate how provenance, XAI, and TAI are interconnected to each other.

3 Provenance, XAI, and TAI: Bibliometric analysis from different aspects

Bibliometric analysis is an effective way to measure the influence of publications in a research area. Our objective behind the bibliometric analysis is to demonstrate evidence of how provenance, XAI, and TAI are interconnected to each other in the publications. To collect the appropriate literature, we compared several databases, such as Google Scholar, PubMed, Web of Science, and Scopus. Although Google Scholar can provide diversified literature, it lacks quality control which makes it inefficient for publication search and analysis. In our work, we decided to focus on only the Scopus database as it provides wide coverage of literature from all major disciplines and all records are organized with good quality measures. A number of terms were used to query the title, abstract, or keywords of publications. As the query script (see below) shows, besides "provenance", we required at least one of the other search terms to be present in the title, abstract, or keywords of a publication. The query was executed in Scopus on August 30th, 2021, and a total of 426 publications between 01/2010 and 12/2020 were found.

Query:

```
( TITLE-ABS-KEY ( machine  AND learning )
OR  TITLE-ABS-KEY ( explainable  AND ai )
OR  TITLE-ABS-KEY ( trustworthy  AND ai )
OR  TITLE-ABS-KEY ( artificial  AND intelligence )
OR  TITLE-ABS-KEY ( explainable  AND artificial  AND intelligence )
AND  TITLE-ABS-KEY ( provenance ) )
AND  PUBYEAR  >  2009  AND  PUBYEAR  <  2021
```

To analyze the results, we used two tools: Bibliometrix and VOS Viewer. Bibliometrix is an open-source tool designed in the R environment for quantitative research, including all the key bibliometric methods of analysis. It allows importing bibliographic data directly from Scopus and other databases. Besides the general bibliometric analysis functions, other measures such as co-citation, coupling, and co-word analysis are also enabled (Aria and Cuccurullo, 2017). VOS Viewer is a software tool for constructing and visualizing bibliometric networks such as authors, journals, and/or individual publications. More sophisticated conditions such as co-occurrences of words or co-citation based on authors can also be used in the network construction (Eck et al., 2010). Below is a list of results generated in our analysis to the 426 publications found on Scopus.

Analysis by timeline of publications: The line graph in Figure 4 shows the number of publications per year from 2010 to 2020. The interesting pattern is an exponential growth in publications from 2016. It shows that the studies related to XAI, TAI, and provenance have received increasing attention in the past four years. Figure 5 is a word growth graph, which shows the cumulative appearance of authors' keywords (i.e., keywords given by authors in a publication) over time among the 426 publications. While overall it shows a trend similar to Figure 4, it is noteworthy that artificial intelligence, machine learning, learning systems, and provenance are the words that stand out as the most predominant among all the authors' keywords.

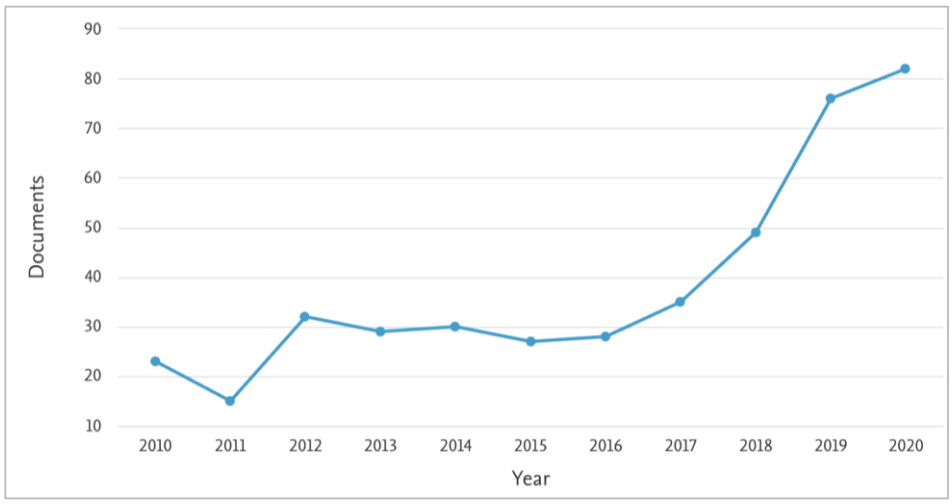

Figure 4: Annual number of publications among the 426 records retrieved from Scopus.

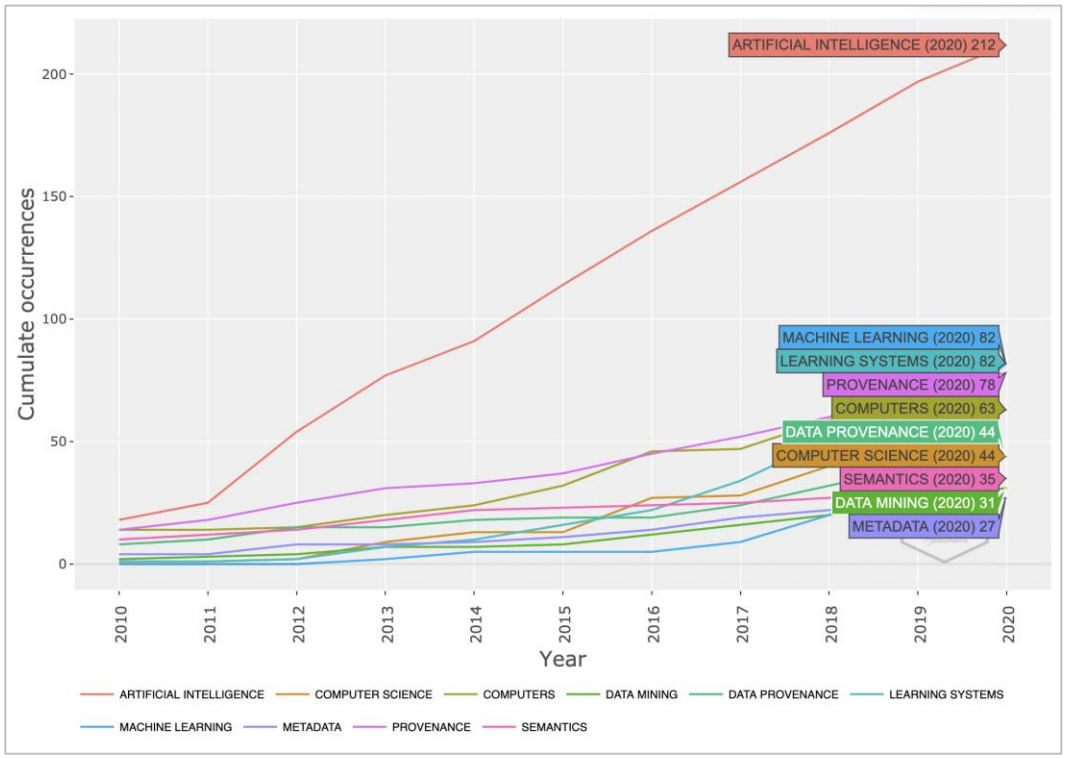

Figure 5: Line graph representing cumulative appearance of word growth among authors' keywords of the 426 publications.

Analysis by subject keywords of references: The references cited by a publication are also a good way to reflect the subject of the publication itself. Keyword Plus collects words or phrases in the titles of a publications references, which provides greater depth and variety for bibliometric analysis (Garfeild, 1990). With Keyword Plus data of the 426 publications retrieved from Scopus, we created a word cloud to visualize the frequency of keywords (Figure 6). The bigger the word or phrase appears in the word cloud, the more often it appears in the Keyword Plus data. Machine learning, learning systems, provenance, data provenance, semantics, and metadata are the most prominent words standing out in the figure.

Figure 6: A word cloud illustrating the most frequent keywords in the Keyword Plus data of the 426 publications.

Analysis by subject area and document type: Another advantage of Scopus data is to show the disciplinary background of the publications. The pie chart in Figure 7 illustrates the proportions of different disciplines among the 426 publications. It is clear that most publications are in the fields of computer science and mathematics. Also, it is interesting to see that about a quarter of the

publications have a background in other disciplines, such as engineering, decision science, and Earth and planetary sciences, which means XAI, TAI, and provenance have also received attention in those disciplines. The donut chart in Figure 8 represents the proportions of document types. Conference papers are more than half and journal articles are about a quarter of the 426 publications.

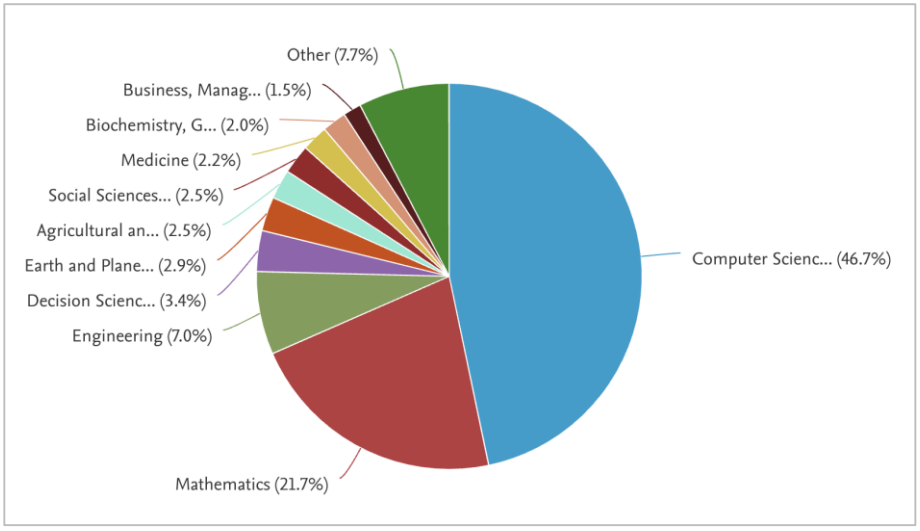

Figure 7: Proportions of disciplines among the 426 publications.

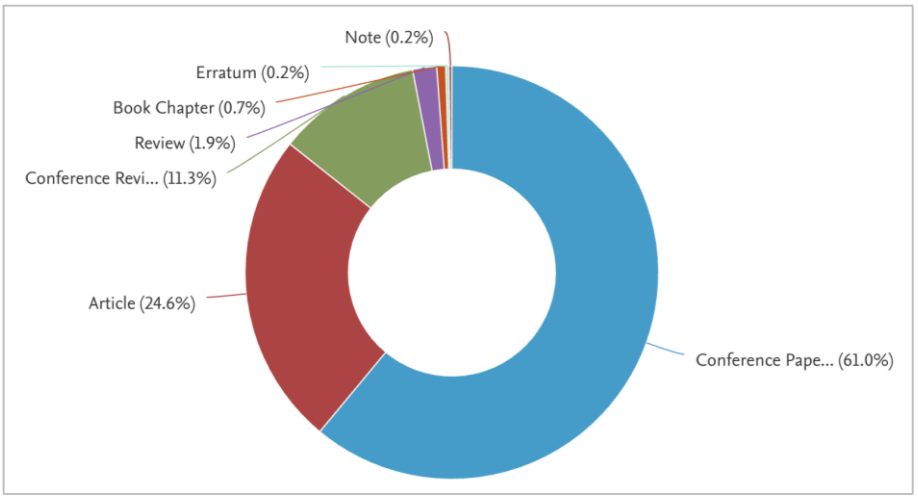

Figure 8: Document types among the 426 publications.

Analysis by co-relationship of authors' keywords: The co-occurrence of authors' keywords shows how different research topics are relevant to each other in a publication. For all the authors' keywords in the 426 publications from Scopus, we first ranked them by frequency of appearance. Then, we took the top 15 keywords in the list and used VOS Viewer to draw a co-occurrence graph (Figure 9). In the figure, the size of each node represents the frequency of appearance of the corresponding keyword. Also, it shows that the 15 keywords are divided into four clusters based on their interconnections, and their frequency of co-occurrence is reflected in the size of lines between the nodes. Among all the 15 keywords and four clusters, provenance and machine learning have the highest appearances. They are closely interconnected with each other and also co-occur with a large number of other keywords.

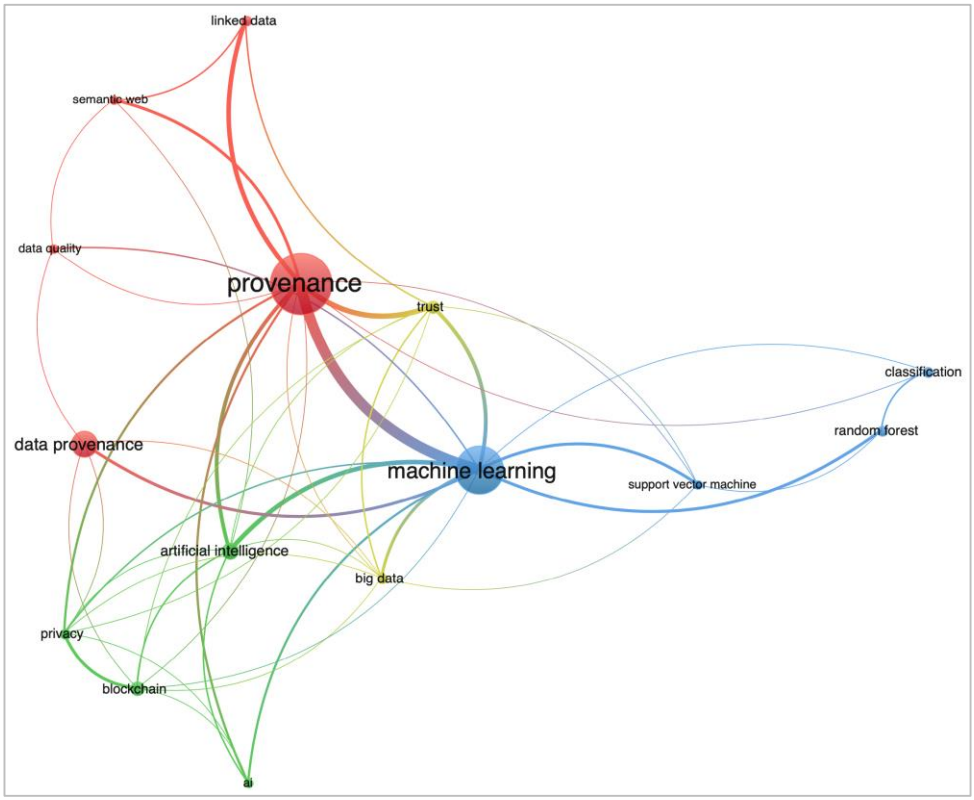

Figure 9: Co-occurrence of authors' keywords among the 426 publications. Here only the top 15 keywords with the highest frequency of appearance are shown.

4 A reflection on the relationship between provenance, XAI, and TAI

4.1 Increasing attention and community works on standards for provenance documentation

The bibliometric analysis in the above section shows an increasing trend of research on provenance, XAI, and TAI. This subsection will incorporate the review of a number of other publications to demonstrate their inter-relationships at a finer scale. Experts and researchers are interested in capturing provenance for several reasons, among which the most important is that well-documented provenance confirms the authenticity of scientific outputs (Moreau et al., 2008). Provenance is the origin or history of something in its literal meaning (Cheney et al., 2009). Some researchers (Jentzsch and Hochgeschwender, 2019) discussed that provenance can be understood as a subset of metadata. We would like to add that provenance not only present the metadata of various objects in a workflow but also the interrelationships between them to show the history of derivation (Ma, 2018). According to PROV Family Documents of the World Wide Web Consortium (W3C), provenance is described as "*information about entities, activities, and people involved in producing a piece of data or thing, which can be used to form assessments about its quality, reliability or trustworthiness*" (Groth and Moreau, 2013; Missier et al., 2013). As such, provenance can answer questions such as how the quality of the data is, what is the data source, when was the data created, what were the steps involved in creating a result, what were the steps in a model used for the data analysis, and who developed and/or ran the workflow (Moreau et al., 2008; Moreau and Groth, 2013).

As AI continues to expand with more diverse information the need of documenting provenance also increases. AI systems need to include provenance as it enables trust and provides users with tools that allow them to access, record, and further investigate resources and steps in a workflow (Chari et al., 2020). The Association for Computing Machinery (ACM) Policy Council set principles for transparency and accountability, in which data provenance is one of the key principles (Garfinkel et al., 2017). Although their comments are on the generic transparency and accountability, their approaches and methods are also insightful for the work of XAI and TAI. Kirkpatrick (2016) stated that regular supervision is necessary for AI-based systems as they can cause harm to many people by generating bias or discriminatory results. Even if the predictions generated by AI/ML models deliver high accuracy, it is crucial to know the very roots before concluding any decision, especially in critical domains such as human activities (Buneman and Tan, 2019; Shaw et al., 2019). Jentzsch and Hochgeschwender (2019) stated that adopting the established methods from the field of provenance to describe ML models will lead to more transparent AI-based systems. A few other researchers also discussed how provenance can increase the reproducibility of ML models (Miles et al., 2007; Davidson and Freire, 2008; Alahmari et al., 2020). Recently, Sarpatwar et al. (2019) described how blockchain allows users to trace the provenance of training models resulting in more transparent and fair AI-based systems. For example, users will be able to discover biases or unclear sourcing of data and see what exactly leads to an action or decision made by AI-based systems. Several other researchers also proposed that provenance is essential to hold AI-based systems to the same standards of accountability as humans (Goodman and Flaxman 2017; Lucero et al., 2018). Based on a review of those publications, in Figure 10 we present the research topics involved in provenance, XAI, and TAI, and illustrate the overlapped parts.

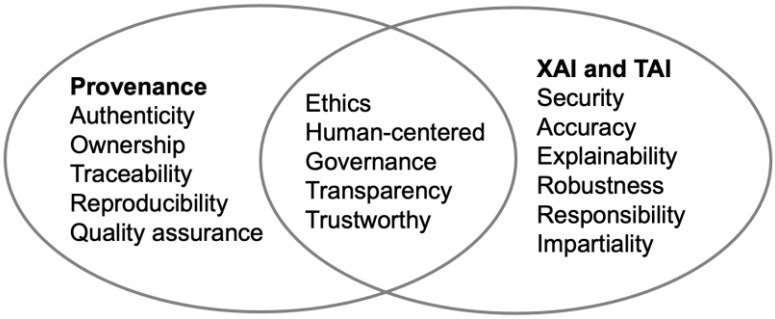

Figure 10: The similarity of topics involved in provenance, XAI, and TAI.

There are many existing models, languages, and tools designed and developed by researchers to enable provenance documentation, and some are developed specifically for AI/ML models. The W3C PROV Ontology (PROV-O) is a representation of the PROV Data Model (PROV-DM) using the Web Ontology Language 2 (OWL2) (Lebo et al., 2013). It allows creating new classes, properties and exchange provenance information generated from different systems. ProvStore is the first online public provenance repository supporting the standards of W3C PROV. It allows users to store, access, integrate, share, organize, visualize, and export provenance documents in various formats, such as PROVN, JSON, Turtle, and XML (Huynh and Moreau, 2014). There are also tools supporting the validation and browsing of provenance documents. ProvValidator is an online tool for validating provenance documents, ensuring that the documents have consistent history and are safe to use for analysis (Moreau et al., 2014). Prov Viewer is a visualization tool that allows users to explore provenance data through zooming, collapsing, filtering to provide different levels of granularity in the analysis (Kohwalter et al., 2016). For workflow platforms and AI/ML models, there are also ongoing activities on specific standards and tools for provenance documentation. The Common Workflow Language (CWL) is a standard designed to provide specifications and semantics for workflows and tools in data-intensive science. The goal is to make

scientific results portable and scalable across software and hardware environments, and thus support reproducibility (Amstutz et al., 2016). OpenML is an online platform that allows machine learning researchers to share the code and results (e.g., model, prediction, and evaluation) and organize it in an effective way for easy access (Vanschoren et al., 2014). ModelDB is an open-source end-to-end system for the management of ML models and has libraries available for Scikit-Learn and Spark ML. It also allows data scientists to perform experiments and build ML models, while the metadata such as pre-processing steps, hyperparameters, quality metrics, and training are automatically captured in the background. ModelDB uses a relational database to store all the extracted metadata and a branching model to track each model's history over time (Vartak et al., 2016).

4.2 Real-world practices of provenance documentation and the support to XAI and TAI

In real-world practice, the scope of provenance differs from user to user and is also dependent on the research needs and technologies used (cf. Simmhan et al., 2005; Buneman et al., 2008; Cheney et al., 2009). To formalized provenance documentation, Groth et al. (2012) outlined the characteristics of the provenance model into several categories, such as content, management, and use. The purpose is to support engineers to categorize the components and dimensions according to the functionality they are involved in. The W3C PROV is a set of documents that defines various aspects necessary to achieve, exchange, and make use of provenance information amongst diverse environments (Groth and Moreau, 2013). For example, the PROV-DM is structured in six components: 1) entities and activities, and the time at which they were created, used, or ended, 2) derivations of entities from other entities, 3) agents bearing responsibility for entities that were generated and activities that happened, 4) a notion of the bundle as a mechanism to support the

provenance of provenance, 5) properties to link entities that refer to the same thing, and 6) collections forming a logical structure for its members (Moreau and Missier, 2013). Those models, categories, and guidelines are further adapted to match needs in real-world applications. For instance, Branco and Moreau (2006) attempted to build a large-scale provenance model for an eScience experiment enabling provenance to be made available as metadata. Pimentel et al. (2016) presented a unique approach for analyzing and tracking provenance collected from scripts. This tool helps scientists record, reproduce, and compare all information and supports decision-making. Huynh et al. (2018) proposed a provenance network analysis method by applying ML techniques on the network metrics to generate provenance information automatically from application data/logs. To provide sufficient information on the decisions made by AI-based systems to the end-users, Jaigirdar (2020) proposed a six-W framework (which, what, who, where, when, and why).

There have been many successful applications of provenance documentation in recent years, and some of them show good performance with AI/ML models in workflow platforms. Renku is an open online platform that can track every version of data, code, and results, and help researchers evaluate, reproduce, and reuse data and algorithms (Krieger et al., 2021). WholeTale is a similar platform that enables reproducibility by allowing researchers to capture and share data, code, and workflow environment in research (Brinckman et al., 2019). Tilmes et al. (2013) and Ma et al. (2014) adapted PROV-O in an ontology to capture provenance of workflows in global change research. Based on those earlier works of provenance documentation, Ma et al. (2017) developed an experiment to capture fine-granular provenance of workflows in Jupyter. Schelter et al. (2017) proposed a lightweight system that allows storage, extraction, and management of

provenance and metadata from ML experiments. Dataset, models, predictions, evaluations, hyperparameters of the models, schemas of the dataset, and layout of the deep neural network are some of the common artifacts that can be achieved. Spinner et al. (2019) designed a visual analytics system named "exlpAIner" which allows users to understand all steps of an ML model, diagnose the limitation using XAI methods, and then refine and optimize the model. Agu et al. (2019) developed a guideline provenance ontology (G-Prov), with the intent to represent provenance of treatments at different granularity levels and share the information with healthcare practitioners. Provenance of scientific workflows has been a long-term concern in research (Davidson et al., 2008). Recently, with the wide usage of Jupyter and RMarkdown in different scientific disciplines, there has also been solid progress on provenance documentation in workflow platforms. For instance, Samuel (2019) designed a tool named ProvBook, which captures and stores the provenance of a notebook in Jupyter and allows users to compare results. ProvPy is a Python library with an implementation of the W3C PROV-DM. It allows to import and export of provenance information in different formats, such as PROV-JSON and PROV-XML (Huynh, 2020).

Some recent projects also leverage the technical advances in semantics, data visualization and cloud computing. For example, MetaClip (METAdata for CLImate Products) (Bedia et al., 2019) develops vocabularies and an R package to capture the provenance of climate research in PROV-O format. The provenance is recorded in JSON-LD format and appended inside the image file of a climate research output. Then, an interactive web portal can load the image and then read and visualize the provenance information into a graph. The nodes and edges in the graph are interactive, where an end user can click and browse the detailed attributes. Another example is Geoweaver

(Sun et al., 2020, 2022), which is an open-source and cloud-based application that allows AI practitioners in earth science to integrate, write, and share workflows. In the cloud-based environment, other users can easily find and trace shared workflows of interest, and replicate the code in their own work.

5 A vision on the trends of provenance, XAI, and TAI in the next decade

It is evident that provenance can help us address issues associated with transparency, explainability, accountability, and authenticity in XAI and TAI. The above bibliometric analysis and reflection highlighted many existing studies, and we believe there will be more advancement in the joint research of XAI, TAI, and provenance in the coming years. Below is a list of our thoughts on future work.

Although AI/ML models have made profound advances, many of them are still deficient in preventing biased and discriminative results. Biases might be caused by many reasons, such as incomplete data, data labelling, adversarial manipulation, missed steps in an ML model, or a workflow guided by a bad hypothesis. Adapting provenance methods will lead to more traceability and transparency of AI applications. A comprehensive description of methods, models, algorithms, and data should be recorded with the aim that they can be further reviewed. Rigorous validation and testing should be done on AI/ML models, and those test results should also be well documented. These steps in provenance documentation can help researchers build explainable and trustworthy systems. Even though documented provenance cannot immediately determine the cause of a bias or error, the complete information can support researchers in tracing all components in the workflow to find the likely cause.

As data are the primary source for any results generated by an AI-based system, studies of XAI and TAI can benefit from many existing mature technologies of metadata and data provenance. Data are suspect when the origin cannot be verified. If a company is using data that are not traceable but concluding an important decision, then this decision is not reliable and will raise concerns amongst users. Provenance provides the flexibility of documenting data at every single step in a data science workflow, ranging from data collection, data cleansing, data analysis, derived data, to the final result. The documented data provenance will be a solid component for XAI and TAI in AI-based systems.

The granularity of provenance (i.e., level of details) depends on the real-world needs. It is crucial to understand that different stakeholders have different requirements on the details of provenance in AI/ML models. Not all people are interested in detailed workflow documentation, while some critical domains such as healthcare, government, and criminal justice require diligent information as the results generated by AI/ML models can have a serious impact on human life, environment, and/or policy making. For AI-based systems, there should be a detailed user survey to clarify the needs of stakeholders before the functions for provenance documentation are developed.

More automated technologies and tools should be developed for recording and sharing provenance information of AI-based systems. We need efficient tools to document provenance and a better-digitized environment to archive, share, and distribute the provenance information to a broad community. Those tools will document the provenance in standard structures and make the information accessible and queryable. In particular, we hope packages can be used for popular

workflow platforms such as Jupyter and RMarkdown to automatically document provenance. Several recent studies mentioned in Section 4 have already made solid progress in that direction. Once those packages are in place, there can be a lot of adoptions and adaptations in various scientific domains.

Moreover, we need to understand XAI and TAI as a socio-technical issue, and we need a comprehensive approach to tackle the issue from both social and technical aspects. The GDPR (General Data Protection Regulation) released by the European Parliament is a good example to help understand this topic. GDPR introduces the standardized data protection law, aiming to create consistent protection of users' data. It states that the data cannot be used without user consent. To assist the implementation of this regulation, provenance information can be used to track down all the activities, which can help to clarify if the data are used in the right way or not. In the world of AI, more work is required to increase awareness and fully establish users' rights and obligations on their data.

6 Conclusions

The need for explainability in AI/ML models has attracted great attention in recent years. However, it is not sufficient to explain AI/ML models using post-hoc explanations alone. Provenance documentation is one of the means to accomplish transparency, traceability, explainability, and reproducibility in AI-based systems. This study presented a systematic literature review of recent work and advances in the field of XAI, TAI, and provenance. First, we provided the fundamental concepts of XAI and TAI and listed the latest discussions on these topics. Second, we analyzed the inter-relationships between XAI, TAI, and provenance through a bibliometric analysis. We

specified how provenance documentation plays a crucial role in building explainability and trustworthiness in AI-based systems, and briefly introduced a few tools and platforms such as Renku, WholeTale, MetaClip, and Geoweaver. Third, we presented a vision on the trends of research on XAI, TAI, and provenance in the next decade. We hope this literature analysis highlights the importance of provenance in AI-based systems and encourages AI practitioners/researchers to start documenting provenance. We expect to see more AI/ML models become explainable, providing enough details to the end-user, and we believe that provenance documentation will be one of the significant approaches to accomplish that.

Acknowledgments

The work was supported by the National Science Foundation under Grants No. 2019609 and the National Aeronautics and Space Administration under Grant No. 80NSSC21M0028. We thank three anonymous reviewers for their constructive comments and suggestions on an earlier version of this paper.

Author Contribution Statement

AK and XM proposed the topic for the literature review and designed the framework. AK conducted the literature review and wrote the first draft. All co-authors contributed to the discussion and revision of the manuscript.

References

Adadi, A. and Berrada, M., 2018. Peeking inside the black box: a survey on explainable artificial intelligence (XAI). IEEE Access, 6, pp.52138-52160.

Agu, N.N., Keshan, N., Chari, S., Seneviratne, O., McCusker, J.P. and McGuinness, D.L., 2019. G-PROV: Provenance Management for Clinical Practice Guidelines. In: Proceedings of the Semantic Web Solutions for Large-scale Biomedical Data Analytics Workshop at the 2019 International Semantic Web Conference, Auckland, New Zealand, pp. 68-75.

AI HLEG (High-Level Expert Group on AI), 2019. Ethics guidelines for trustworthy AI. European Commission, Brussels, 39pp. Available: https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai. Accessed on: September 21, 2021.

Alahmari, S.S., Goldgof, D.B., Mouton, P.R. and Hall, L.O., 2020. Challenges for the Repeatability of Deep Learning Models. IEEE Access, 8, pp.211860-211868.

Amalina, F., Hashem, I.A.T., Azizul, Z.H., Fong, A.T., Firdaus, A., Imran, M. and Anuar, N.B., 2019. Blending Big Data Analytics: Review on Challenges and a Recent Study. IEEE Access, 8, pp.3629-3645.

Amstutz, P., Crusoe, M.R., Tijanić, N., Chapman, B., Chilton, J., Heuer, M., Kartashov, A., Leehr, D., Ménager, H., Nedeljkovich, M. and Scales, M., 2016. Common workflow language, V1.0. Figshare. Available: https://doi.org/10.6084/m9.figshare.3115156.v2.

Aria, M. and Cuccurullo, C., 2017. bibliometrix: An R-tool for comprehensive science mapping analysis. Journal of Informetrics, 11(4), pp.959-975.

Arrieta, A.B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., García, S., Gil-López, S., Molina, D., Benjamins, R. and Chatila, R., 2020. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. Information Fusion, 58, pp.82-115.

Bedia, J., San-Martín, D., Iturbide, M., Herrera, S., Manzanas, R. and Gutiérrez, J.M., 2019. The METACLIP Semantic Provenance Framework for Climate Products. Environmental Modelling & Software, 119, pp. 445-457.

Belle, V. and Papantonis, I., 2021. Principles and Practice of Explainable Machine Learning. Frontiers in Big Data, 4, 25pp. Available: http://dx.doi.org/10.3389/fdata.2021.688969.

Boudette, N.E., 2016. Autopilot Cited in Death of Chinese Tesla Driver. The New York Times, 14 September. https://www.nytimes.com/2016/09/15/business/fatal-tesla-crash-in-china-involved-autopilot-government-tv-says.html. Accessed on: September 21, 2021.

Branco, M. and Moreau, L., 2006. Enabling provenance on large scale e-science applications. In: Proceedings of the 2006 International Provenance and Annotation Workshop, Chicago, IL, USA, pp. 55-63.

Brinckman, A., Chard, K., Gaffney, N., Hategan, M., Jones, M.B., Kowalik, K., Kulasekaran, S., Ludäscher, B., Mecum, B.D., Nabrzyski, J. and Stodden, V., 2019. Computing environments for reproducibility: Capturing the "Whole Tale". Future Generation Computer Systems, 94, pp.854-867.

Buneman, P. and Tan, W.C., 2019. Data provenance: What next? ACM SIGMOD Record, 47(3), pp. 5-16.

Buneman, P., Cheney, J., Tan, W.C. and Vansummeren, S., 2008, June. Curated databases. In: Proceedings of the 27th ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems, Vancouver, Canada, pp. 1-12.

Castelvecchi, D., 2016. Can we open the black box of AI? Nature News, 538(7623), pp. 20 - 23.

Chari, S., Gruen, D., Seneviratne, O. and Mcguinness, D., 2020. Foundations of Explainable Knowledge-Enabled Systems. Available at: https://arxiv.org/pdf/2003.07520.pdf

Chen, L., Cruz, A., Ramsey, S., Dickson, C.J., Duca, J.S., Hornak, V., Koes, D.R. and Kurtzman, T., 2019. Hidden bias in the DUD-E dataset leads to misleading performance of deep learning in structure-based virtual screening. PloS One, 14(8), e0220113.

Cheney, J., Chiticariu, L. and Tan, W.C., 2009. Provenance in Databases: Why, How, and Where. Now Publishers Inc. Hanover, MA, USA. 100p. Available: http://dx.doi.org/10.1561/9781601982339.

Dastin, J., 2018. Amazon scraps secret AI recruiting tool that showed bias against women. Reuters. https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G. Accessed on: September 23, 2021.

Davidson, S.B. and Freire, J., 2008. Provenance and scientific workflows: challenges and opportunities. In: Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data, Vancouver, Canada, pp. 1345-1350.

Floridi, L., 2019. Establishing the rules for building trustworthy AI. Nature Machine Intelligence, 1(6), pp. 261-262.

Frost, L., 2019. Explainable AI and other questions where provenance matters, IEEE IoT Newsletter. https://iot.ieee.org/newsletter/january-2019/explainable-ai-and-other-questions-where-provenance-matters. Accessed on: September 21, 2021.

Garfield, E., 1990. KeyWords Plus-ISI's breakthrough retrieval method. 1. Expanding your searching power on current-contents on diskette. Current Contents, 32, pp. 5-9.

Garfinkel, S., Matthews, J., Shapiro, S.S. and Smith, J.M., 2017. Toward algorithmic transparency and accountability. Communications of the ACM, 60(9), pp. 5–5.

Goddard, M., 2017. The EU General Data Protection Regulation (GDPR): European regulation that has a global impact. International Journal of Market Research, 59(6), pp. 703-705.

Goodfellow, I., Bengio, Y. and Courville, A., 2016. Deep Learning. MIT Press, Cambridge, MA, 800p.

Goodman, B. and Flaxman, S., 2017. European Union regulations on algorithmic decision-making and a "right to explanation". AI Magazine, 38(3), pp. 50-57.

Groth, P. and Moreau, L., 2013. An Overview of the PROV Family of Documents, W3C. https://www.w3.org/TR/prov-overview/. Accessed on: September 21, 2021.

Groth, P., Gil, Y., Cheney, J. and Miles, S., 2012. Requirements for provenance on the web. International Journal of Digital Curation, 7(1), pp. 39-56.

Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F. and Pedreschi, D., 2018. A survey of methods for explaining black box models. ACM Computing Surveys, 51(5), pp. 1-42.

Gunning, D. and Aha, D., 2019. DARPA's explainable artificial intelligence (XAI) program. AI Magazine, 40(2), pp. 44-58.

Gunning, D., 2019. DARPA's Explainable artificial intelligence (XAI) program. In: Proceedings of the 24th International Conference on Intellegent Users interface. Marina del Ray, CA, USA. pp. ii-ii.

Holzinger, A., Biemann, C., Pattichis, C.S. and Kell, D.B., 2017. What do we need to build explainable AI systems for the medical domain? 28pp. Available: https://arxiv.org/abs/1712.09923.

Huynh, T.D. and Moreau, L., 2014. ProvStore: a public provenance repository. In: Proceedings of the 2014 International Provenance and Annotation Workshop, Cologne, Germany, pp. 275-277.

Huynh, T.D., 2020. Prov 2.0.0 Python Package. https://pypi.org/project/prov/. Accessed on: September 21, 2021.

Huynh, T.D., Ebden, M., Fischer, J., Roberts, S. and Moreau, L., 2018. Provenance Network Analytics. Data Mining and Knowledge Discovery, 32(3), pp. 708-735.

Jaigirdar, F.T., Rudolph, C. and Bain, C., 2019. Can I trust the data I see? A Physician's concern on medical data in IoT health architectures. In: Proceedings of the Australasian Computer Science Week Multiconference, Sydney, Australia. pp. 1-10.

Jaigirdar, F.T., Rudolph, C., Oliver, G., Watts, D. and Bain, C., 2020. What Information is Required for Explainable AI?: A Provenance-based Research Agenda and Future Challenges. In: Proceedings of the IEEE 6th International Conference on Collaboration and Internet Computing (CIC). Atlanta, GA, USA. pp. 177-183.

Jain, S., Luthra, M., Sharma, S. and Fatima, M., 2020. Trustworthiness of Artificial Intelligence. In: Proceedings of the 6th International Conference on Advanced Computing and Communication Systems (ICACCS). Coimbatore, India. pp. 907-912.

Jentzsch, S.F. and Hochgeschwender, N., 2019. Don't Forget Your Roots! Using Provenance Data for Transparent and Explainable Development of Machine Learning Models. In: Proceedings of the 34th IEEE/ACM International Conference on Automated Software Engineering Workshop (ASEW). San Diego, CA, USA. pp. 37-40.

Kirkpatrick, K., 2016. Battling algorithmic bias. Communications of the ACM, 59(10), pp. 16–17.

Kohwalter, T., Oliveira, T., Freire, J., Clua, E. and Murta, L., 2016. Prov viewer: A graph-based visualization tool for interactive exploration of provenance data. In: Proceedings of the International Provenance and Annotation Workshop. McLean, VA, USA. pp. 71-82.

Krieger, L., Nijzink, R., Thakur, G., Ramakrishnan, C., Roskar, R. and Schymanski, S., 2021. Repeatable and reproducible workflows using the RENKU open science platform. In: Proceedings of 2021 EGU General Assembly, Virtual Meeting. Abstract EGU21-7655.

Kumar, A., McCann, R., Naughton, J. and Patel, J.M., 2016. Model selection management systems: The next frontier of advanced analytics. ACM SIGMOD Record, 44(4), pp.17-22.

Lebo, T., Sahoo, S., McGuinness, D., Belhajjame, K., Cheney, J., Corsar, D., Garijo, D., Soiland-Reyes, S., Zednik, S., and Zhao, J., 2013. PROV-O: The PROV Ontology. (W3C Recommendation). World Wide Web Consortium. http://www.w3.org/TR/2013/REC-prov-o-20130430/. Accessed on: September 21, 2021.

Liu, S., Wang, X., Liu, M. and Zhu, J., 2017. Towards better analysis of machine learning models: A visual analytics perspective. Visual Informatics, 1(1), pp. 48-56.

Lucero, C., Coronado, B., Hui, O. and Lange, D.S., 2018. Exploring explainable artificial intelligence and autonomy through provenance. In: Proceedings of the 2nd Workshop on Explainable Artificial Intellgence. Stockholm, Sweden. pp. 85 -89.

Lundberg, S.M. and Lee, S.I., 2017, December. A unified approach to interpreting model predictions. In Proceedings of the 31st International Conference on Neural Information Processing Systems, Long Beach, CA, USA, pp. 4768-4777.

Lundberg, S.M., Erion, G., Chen, H., DeGrave, A., Prutkin, J.M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N. and Lee, S.I., 2020. From local explanations to global understanding with explainable AI for trees. Nature Machine Intelligence, 2(1), pp.56-67.

Ma, X., 2018. Metadata. In: Schintler, L.A., McNeely, C.L. (eds.) Encyclopedia of Big Data. Springer, Cham, Switzerland. 5pp. doi:10.1007/978-3-319-32001-4_135-1.

Ma, X., Beaulieu, S.E., Fu, L., Fox, P., Di Stefano, M., West, P., 2017. Documenting Provenance for Reproducible Marine Ecosystem Assessment in Open Science. In: Diviacoo, P., Glaves, H.M., Leadbetter, A. (eds.) Oceanographic and Marine Cross-Domain Data Management for Sustainable Development. IGI Global, Hershey, PA, USA. pp. 100-126.

Ma, X., Zheng, J.G., Goldstein, J., Zednik, S., Fu, L., Duggan, B., Aulenbach, S., West, P., Tilmes, C., Fox, P., 2014. Ontology engineering in provenance enablement for the National Climate Assessment. Environmental Modelling & Software 61, pp. 191-205.

McCausland, P., 2019. Self-driving Uber car that hit and killed woman did not recognize that pedestrians jaywalk, NBC News, 9 November. https://www.nbcnews.com/tech/tech-news/self-driving-uber-car-hit-killed-woman-did-not-recognize-n1079281. Accessed on: September 21, 2021.

McFarland, M., 2018. Uber shuts down self-driving operations in Arizona. https://money.cnn.com/2018/05/23/technology/uber-arizona-self-driving/index.html. Accessed on: September 21, 2021.

MEAEF (Ministry of Economic Affairs and Employment of Finland), 2019. Leading the Way into the Era of Artificial Intelligence: Final Report of Finland's Artificial Intelligence Program 2019. Ministry of Economic Affairs and Employment of Finland. 133pp. Available: http://urn.fi/URN:ISBN:978-952-327-437-2.

Miles, S., Wong, S.C., Fang, W., Groth, P., Zauner, K.P. and Moreau, L., 2007. Provenance-based validation of e-science experiments. Journal of Web Semantics, 5(1), pp. 28-38.

Miller, T., 2019. Explanation in artificial intelligence: Insights from the social sciences. Artificial Intelligence, 267, pp. 1-38.

Missier, P., Belhajjame, K. and Cheney, J., 2013. The W3C PROV family of specifications for modelling provenance metadata. In: Proceedings of the 16th International Conference on Extending Database Technology, Genoa, Italy. pp. 773-776.

Montavon, G., Lapuschkin, S., Binder, A., Samek, W. and Müller, K.R., 2017. Explaining nonlinear classification decisions with deep taylor decomposition. Pattern Recognition, 65, pp. 211-222.

Moreau, L. and Groth, P., 2013. Provenance: An introduction to PROV. Synthesis Lectures on the Semantic Web: Theory and Technology, 3(4), pp. 1-129.

Moreau, L. and Missier, P., 2013. PROV DM: The PROV Data Model, W3C. https://www.w3.org/TR/2013/REC-prov-dm-20130430/. Accessed on: September 21, 2021.

Moreau, L., Groth, P., Miles, S., Vazquez-Salceda, J., Ibbotson, J., Jiang, S., Munroe, S., Rana, O., Schreiber, A., Tan, V. and Varga, L., 2008. The provenance of electronic data. Communications of the ACM, 51(4), pp.52-58.

Moreau, L., Huynh, T.D. and Michaelides, D., 2014. An online validator for provenance: Algorithmic design, testing, and API. In: Proceedings of the International Conference on Fundamental Approaches to Software Engineering, Grenoble, France. pp. 291-305.

Moreau, L., Ludäscher, B., Altintas, I., Barga, R. S., Bowers, S., Callahan, S., Chin, G., Clifford, B., Cohen, S., Cohen-Boulakia, S., Davidson, S., Deelman, E., Digiampietri, L., Foster, I., Freire, J., Frew, J., Futrelle, J., Gibson, T., Gil, Y., Goble, C., Golbeck, J., Groth, P., Holland, D. A., Jiang, S., Kim, J., Koop, D., Krenek, A., McPhillips, T., Mehta, G., Miles, S., Metzger, D., Munroe, S., Myers, J., Plale, B., Podhorszki, N., Ratnakar, V., Santos, E., Scheidegger, C., Schuchardt, K., Seltzer, M., Simmhan, Y. L., Silva, C., Slaughter, P., Stephan, E., Stevens, R.,

Turi, D., Vo, H., Wilde, M., Zhao, J. and Zhao, Y., 2008. Special Issue: The First Provenance Challenge. Concurrency and Computation: Practice and Experience, 20(5), pp. 409–418.

Murdoch, W.J., Singh, C., Kumbier, K., Abbasi-Asl, R. and Yu, B., 2019. Interpretable machine learning: definitions, methods, and applications. pp. 1-11. Available: https://arxiv.org/abs/1901.04592.

NAS (National Academies of Sciences), 2018. The Frontiers of Machine Learning: 2017 Raymond and Beverly Sackler U.S.-U.K. Scientific Forum. The National Academies Press, Washington, DC, 32pp. DOI: 10.17226/25021.

Osoba, O.A. and Welser IV, W., 2017. An Intelligence in Our Image: The Risks of Bias and Errors in Artificial Intelligence. RAND Corporation, Santa Monica, CA, USA, 34pp.

Pimentel, J.F., Freire, J., Braganholo, V. and Murta, L., 2016. Tracking and analyzing the evolution of provenance from scripts. In: Proceedings of the 6th International Provenance and Annotation Workshop, McLean, VA, USA. pp. 16-28.

Rakova, B., Chowdhury, R. and Yang, J., 2020. Assessing the intersection of organizational structure and FAT* efforts within industry: implications tutorial. In: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, Barcelona, Spain. pp. 697-697.

Ribeiro, M.T., Singh, S. and Guestrin, C., 2016. Why should i trust you? Explaining the predictions of any classifier. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA. pp. 1135-1144.

Roberts, H., Cowls, J., Morley, J., Taddeo, M., Wang, V. and Floridi, L., 2021. The Chinese approach to artificial intelligence: an analysis of policy, ethics, and regulation. AI & Society, 36(1), pp. 59-77.

Ross, C. and Swetlitz, I., 2018. IBM's Watson supercomputer recommended 'unsafe and incorrect' cancer treatments, internal documents show. https://www.statnews.com/wp-content/uploads/2018/09/IBMs-Watson-recommended-unsafe-and-incorrect-cancer-treatments-STAT.pdf Accessed on: September 21, 2021.

Rudin, C. and Radin, J., 2019. Why Are We Using Black Box Models in AI When We Don't Need To? A Lesson From An Explainable AI Competition. Harvard Data Science Review, 1(2), doi:10.1162/99608f92.5a8a3a3d.

Rudin, C., 2018. Please stop explaining black box models for high stakes decisions. In: Proceedings of the 32nd Conference of Neural Information Processing Systems (NIPS), Workshop on Critiquing and Correcting Trends Machine Learning, Montreal, Canada, 20pp. Available: https://arxiv.org/abs/1811.10154.

Rudin, C., 2019. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature Machine Intelligence, 1(5), pp. 206-215.

Samuel, S., 2019. A provenance-based semantic approach to support understandability, reproducibility, and reuse of scientific experiments. Doctoral dissertation, Friedrich-Schiller-Universität Jena, 241pp.

Samuel, S., Löffler, F. and König-Ries, B., 2020. Machine learning pipelines: provenance, reproducibility and FAIR data principles. In: Proceedings of the 2021 International Provenance and Annotation Workshop, Charlotte, NC. Available at: https://arxiv.org/abs/2006.12117.

Sarpatwar, K., Vaculin, R., Min, H., Su, G., Heath, T., Ganapavarapu, G. and Dillenberger, D., 2019. Towards Enabling Trusted Artificial Intelligence via Blockchain. In: Calo, S., Bertino, E., Verma, D. (eds.) Policy-Based Autonomic Data *Governance*. Springer, Cham, pp. 137-153.

Schelter, S., Boese, J.H., Kirschnick, J., Klein, T. and Seufert, S., 2017. Automatically tracking metadata and provenance of machine learning experiments. In: Proceedings of Machine Learning Systems Workshop at the 31st Conference on Neural Information Processing Systems, Long Beach, CA, USA, pp. 27-29.

Shaw, J., Rudzicz, F., Jamieson, T. and Goldfarb, A., 2019. Artificial intelligence and the implementation challenge. Journal of Medical Internet Research, 21(7), e13659.

Simmhan, Y.L., Plale, B. and Gannon, D., 2005. A survey of data provenance in e-science. ACM SIGMOD Record, 34(3), pp. 31-36.

Singh, J., Cobbe, J. and Norval, C., 2018. Decision provenance: Harnessing Data Flow for Accountable Systems. IEEE Access, 7, pp.6562-6574.

Spinner, T., Schlegel, U., Schäfer, H. and El-Assady, M., 2019. explAIner: A visual analytics framework for interactive and explainable machine learning. IEEE Transactions on Visualization and Computer Graphics, 26(1), pp. 1064-1074.

Sun, Z., Di, L., Burgess, A., Tullis, J.A. and Magill, A.B., 2020. Geoweaver: Advanced cyberinfrastructure for managing hybrid geoscientific AI workflows. ISPRS International Journal of Geo-Information, 9(2), 119. doi: 10.3390/ijgi9020119.

Sun, Z., Sandoval, L., Crystal-Ornelas, R., Mousavi, S.M., Wang, J., Lin, C., Cristea, N., Tong, D., Carande, W.H., Ma, X. and Rao, Y., 2022. A review of Earth Artificial Intelligence. Computers & Geosciences, 159,105034. doi:10.1016/j.cageo.2022.105034.

Sundararajan, M., Taly, A. and Yan, Q., 2017. Axiomatic Attribution for Deep Networks. In: Proceedings of the 34th International Conference on Machine Learning, Sydney, Australia, pp. 3319-3328.

Tan, S., Caruana, R., Hooker, G. and Lou, Y., 2017. Detecting bias in black-box models using transparent model distillation. In: Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society (AIES '18), New Orleans, LA, USA, pp. 303–310.

Tennery, A. and Cherelus, G., 2016. Microsoft's AI Twitter bot goes dark after racist, sexist tweets, Reuters, 24 March. https://www.reuters.com/article/us-microsoft-twitter-bot/microsofts-ai-twitter-bot-goes-dark-after-racist-sexist-tweets-idUSKCN0WQ2LA. Accessed on: September 21, 2021.

Thiebes, S., Lins, S. and Sunyaev, A., 2020. Trustworthy artificial intelligence. Electronic Markets, pp.447–464.

Tilmes, C., Fox, P., Ma, X., McGuinness, D., Privette, A.P., Smith, A., Waple, A., Zednik, S., Zheng, J., 2013. Provenance representation for the National Climate Assessment in the Global Change Information System. IEEE Transactions on Geoscience and Remote Sensing 51 (11), pp. 5160-5168.

Van Eck, N.J. and Waltman, L., 2010. Software survey: VOSviewer, a computer program for bibliometric mapping. Scientometrics, 84(2), pp. 523-538.

Vanschoren, J., Van Rijn, J.N., Bischl, B. and Torgo, L., 2014. OpenML: networked science in machine learning. ACM SIGKDD Explorations Newsletter, 15(2), pp. 49-60.

Vartak, M., Subramanyam, H., Lee, W.E., Viswanathan, S., Husnoo, S., Madden, S. and Zaharia, M., 2016, June. ModelDB: a system for machine learning model management. In: Proceedings of the Workshop on Human-In-the-Loop Data Analytics, San Francisco, CA, USA. pp. 1-3.

Vincent, J., 2016. Twitter taught Microsoft’s AI chatbot to be a racist asshole in less than a day, The Everge, 24 March. https://www.theverge.com/2016/3/24/11297050/tay-microsoft-chatbot-racist. Accessed on: September 21, 2021.

Werder, K. and Ramesh, B., 2021. Establishing Data Provenance for Responsible Artificial Intelligence Systems. ACM Transactions on Management Information Systems, In Press.

White House, 2020. Executive Order on Promoting the Use of Trustworthy Artificial Intelligence in the Federal Government. https://trumpwhitehouse.archives.gov/articles/promoting-use-trustworthy-artificial-intelligence-government/. Accessed on: September 23, 2021.

Wing, J.M., 2020. Ten Research Challenge Areas in Data Science. Harvard Data Science Review, 2(3), doi:10.1162/99608f92.c6577b1f.

Wing, J.M., 2021. Trustworthy AI. Communications of the ACM, 64(10), pp. 1-12.

Wohlin, C., 2014. Guidelines for snowballing in systematic literature studies and a replication in software engineering. In: Proceedings of the 18th International Conference on Evaluation and Assessment in Software Engineering, Karlskrona, Sweden. pp. 1-10.