Establishing Data Provenance for Responsible Artificial Intelligence Systems

Karl Werder[†]
Cologne Institute for Information Systems
University of Cologne
Germany
werder@wiso.uni-koeln.de

Balasubramaniam Ramesh
Computer Information Systems
Georgia State University
USA
bramesh@gsu.edu

Rongen (Sophia) Zhang
Computer Information Systems
Georgia State University
USA
rzhang6@gsu.edu

ABSTRACT

Data provenance, a record that describes the origins and processing of data, offers new promises in the increasingly important role of artificial intelligence (AI)-based systems in guiding human decision making. To avoid disastrous outcomes that can result from bias-laden AI systems, responsible AI builds on four important characteristics: fairness, accountability, transparency, and explainability. To stimulate further research on data provenance that enables responsible AI, this study outlines existing biases and discusses possible implementations of data provenance to mitigate them. We first review biases stemming from the data's origins and pre-processing. We then discuss the current state of practice, the challenges it presents, and corresponding recommendations to address them. We present a summary highlighting how our recommendations can help establish data provenance and thereby mitigate biases stemming from the data's origins and pre-processing to realize responsible AI-based systems. We conclude with a research agenda suggesting further research avenues.

CCS CONCEPTS

•Data Provenance •Artificial Intelligence

KEYWORDS

Data Provenance, Artificial Intelligence, Fairness, Accountability, Transparency, Explainability

1 Introduction

As evidence-based decision making aided by data-driven artificial intelligence (AI) algorithms becomes increasingly common across all sectors of the economy, there is a growing concern among users about whether such algorithms are developed and implemented responsibly. Prior reports have already provided a glimpse into the disastrous effects of inaccurate and bias-laden AI recommendations in high-stakes applications, with examples from the healthcare and legal domains, such as incorrect patient treatment, exacerbated poverty [62], wrongful arrest [33], and unjust criminal sentencing [43]. The heightened awareness of concerns raised in recent movements for social justice has resulted in calls from professional associations [1] and researchers [18,34] for developing approaches that help establish responsible AI.

Rapid innovations in data-generating technologies, such as sensors, social media, and mobile devices, have exacerbated the problems resulting from poor data quality that threaten the development of responsible AI systems. These technologies generate an unprecedented quantity and variety of data. While most applications have benefitted from explosive growth in data availability (in terms of volume, variety, velocity, veracity, etc.), limited attention has been given to data quality [66], in turn undermining the quality of recommendations generated using such data. Motivated by these concerns, this study examines how data provenance can help improve data quality and enhance the fairness, accountability, transparency, and explainability (FATE) of AI-based systems. We argue that data provenance—a record that describes the origins and processing of data [9]—can help assess and improve the FATE of recommendations provided by AI algorithms and thus instill trust in them. Trust is enhanced by the capability to describe and follow the life of data (i.e., their origins, processing, and use) in both forward and backward directions [75]. The importance of provenance has long been recognized [14] in the pharmaceutical, food, and fashion industries. It helps establish a product's origins and influences consumers' decisions about purchase and use.

Responsible AI is essentially related to a broad discourse, AI ethics, which has received significant attention among researchers in recent years. Scholars have identified different high-level ethical principles that should govern the development of AI systems [25,48,97]. While no universal consensus exists, fairness, accountability, and transparency [48] have received significant attention in this research community [27]. Simultaneously, research related to explainable AI has emerged [39], with recent discussions on its capability to bridge the gap between technical and ethical considerations [64]. AI explainability gives users and experts the ability to investigate and understand the inner workings of AI,

[†]Karl Werder, Cologne Institute for Information Systems, University of Cologne, Pohligstr.1, 50969 Cologne, Germany

allowing them to identify potential biases. Bridging these two perspectives, we focus on four important and related characteristics of responsible AI—FATE. While there is ongoing research on other AI-based system characteristics, such as privacy and agency, we focus on how FATE can help organizations identify and mitigate the negative influences of biases within their data. We discuss how potential conflicts among different FATE characteristics emerge, how organizations can manage them, and where more research is needed.

Most current researchers and practitioners in the field of responsible AI have emphasized the quality of algorithms. However, an algorithm's recommendations or outputs also depend heavily on representations, structures, and data quality, which serve as the inputs. In this study, we focus on data provenance, an important aspect of data quality, in the development of responsible AI systems [13]. For example, data provenance can help uncover data quality concerns related to labor-intensive data labeling, which is often performed by unqualified workers [7] and otherwise remains concealed. This is particularly alarming, as the recommendations or outputs of AI algorithms are often used as inputs for other AI algorithms [53], further exacerbating the problem. For example, the classification of a radiology scan by an algorithm as benign or malignant may be used as an input for another algorithm that is used to create a risk score for patient readmission. In such situations, data provenance can help identify the causes of the AI algorithm's poor performance, improve interpretability, or uncover that its seemingly acceptable performance was achieved for invalid reasons (e.g., when identifying a malignant tumor, the system was learning from the circle made by the radiologist on the scan rather than the data from the scan itself). By illuminating the origin and processing of the data [14], data provenance can mitigate these shortcomings and facilitate FATE assessments (see Table 1).

The lack of data provenance is a serious concern in AI-based systems that are used to inform critical decisions. While the establishment of data provenance may increase short-term costs for organizations, it can provide long-term benefits by instilling trust in the implemented system and its recommendations. Specifically, our study addresses the following question: ***How does data provenance affect the four interrelated characteristics of responsible AI: fairness, accountability, transparency, and explainability?***

The paper analyzes biases related to origins and pre-processing of data, discusses the current state of practice and attendant challenges, and presents recommendations for addressing them. Our recommendations are intended to help establish data provenance and mitigate biases stemming from the data's origins and pre-processing to realize responsible AI-based systems.

Table 1 – Overview of FATE characteristics and examples

Characteristic	Description	Example
Fairness	AI-based systems may introduce discrimination because of imbalanced data [4]. The data used in training AI-based systems often reflect the discriminations existing in our society, which, in turn, lead to algorithmic bias [4].	Training the system using only medical records from male patients can lead to discrimination against female patients.
Accountability	Because of the increasing complexity of AI-based systems, it is difficult for a user to judge who is accountable for the results [49]. The individual services provided by AI algorithms are integrated into larger systems [19], further exacerbating opaqueness and ambiguity about ownership.	When an AI-based system trained on photos depicting cancer on the epidermis (outer skin layer) is integrated into a larger system, it may also be inappropriately used on data from subcutaneous tissue (inner skin layer). It becomes unclear who is accountable for the resulting incorrect recommendation.
Transparency	An often-cited limitation of AI-based systems is their black box nature [2]. However, to understand the quality of recommendations and training data adequacy, we need transparency.	The pharma industry has well-established practices for providing easy access to relevant information about drugs (either in the product package itself or in the accompanying documents), whereas AI systems seldom provide relevant information about the data used in developing recommendations.
Explainability	A lack of explainability of AI prediction outcomes can be caused by the black box nature of algorithms, which can lead to negligence of the inaccuracies and biases in data. Yet, understanding a prediction is an important aspect of their acceptance [84].	Evidence-based medicine rests on high standards of explainability of both algorithms and data, as medical decision making requires a sound understanding of the underlying disease mechanisms and treatments [88]. The lack of this understanding undermines the implementation of AI in healthcare [81].

In the following sections, we review key biases, such as systematic distortions [3], resulting from the failure to adopt appropriate data provenance practices in the development and implementation of AI-based systems. We also provide three key recommendations for establishing data provenance to enhance the FATE of AI-based systems. We propose a data provenance framework for responsible AI and discuss exemplary cases for its application. Before concluding, we present future research directions.

2 Sources of Data Biases in AI-based Systems

In contrast to the majority of existing research, which has focused on biases resulting from algorithms (e.g., [28,35]), we concentrate on the origins of the data and the data pre-processing rather than on the algorithm that uses the data as inputs. Data sources are often where the original data were collected to train and build AI-based systems. After data collection, data pre-processing [30], which commonly includes data preparation, integration, cleaning, normalization, and transformation, can also introduce biases [96]. We identify five categories of potential biases that may originate from data sources and five categories of biases that may be introduced during data pre-processing. For example, the data themselves might be subject to bias in the ways in which they are sampled or measured. Each bias has different implications for the FATE characteristics of AI-based systems.

Table 2 – Summary of the effect of data biases on responsible AI

Origins	Bias	Fairness	Accountability	Transparency	Explainability
Data Source	Population data	X		X	X
	Measurement error	X			X
	Data quality chasm	X	X	X	
	Data repurposing	X	X		X
	Data augmentation	X	X	X	
Pre-Processing	Dataset shifts	X		X	
	Opaque pre-processing			X	X
	Data labeling	X	X	X	X
	Adversarial manipulations	X		X	X
	Transfer learning	X	X	X	

2.1 Biases from the Data's Origins

Below, we identify five key instances in which biases arise in the data sources: population data, measurement error, data quality chasm, data repurposing, and data augmentation. We describe their implications regarding the FATE characteristics (see Table 2 for a summary).

Population data. In every data science project, sampling the right data to ensure representativeness is important [33]. However, to develop and implement powerful AI-based systems, developers often rely on access to unique data. For example, data provided through projects, such as BigMedilytics, comprise the medical records of more than 11 million patients from eight countries. The retraining or recalibration of AI-based systems developed with such unique data to other contexts for the same purpose requires additional data that are representative of the new context.

However, AI-based systems are often applied in new contexts without retraining or recalibration because of the significant challenges involved in collecting the necessary additional data. For example, when an algorithm is trained with data from one population but is used to develop predictions on another population, any differences in the frequency and nature of events in these datasets will result in poor performance [19]. When the data collection mechanisms impose selection bias or fail to recognize the mismatch between the training data and the target population, the transparency of the data's origins is affected. In addition, spurious correlations and shortcut learning (i.e., decision rules that work well based on the training data because of spurious phenomena [32]) of the AI system will lead to unreliable and unfair recommendations [20] that will undermine possible explanations.

Measurement error. Every study and every measurement instrument, however well designed, still generates some errors [72]. Many AI applications in domains such as medicine or business rely heavily on Bayesian statistics, as the results are always subject to probabilities. Data pre-processing and the use of another algorithm's predictions as an input could further compound this issue because of the propagation of uncertainties or prior probabilities [61].

However, in AI systems, the uncertainty of the input variables resulting either from the measurement itself or from pre-processing is often neglected. An AI-based system trained with such data without a particular focus on and caution about potential errors can result in a poorly performing model. Consequently, the precision of an AI-based system might be overestimated, as the AI system learns to fit against the error. The resulting recommendations would be at least distorted if not incorrect, leading to problematic outcomes. If the system provides corresponding explanations, a user can identify these inadequacies and correct them [19].

Data quality chasm. Another challenge is the lack of data with adequate quality in settings where the AI system is used [61]. While the data may look homogeneous at the surface level, a more careful evaluation can suggest otherwise. For example, an AI algorithm may achieve superior prediction quality because of its access to state-of-the-art computed tomography (CT) scans. If CT scans from older equipment that generates lower-quality scans are used to retrain the AI-based system, the recommendations are likely to be inaccurate.

In contrast to the measurement error, in which the system has learned to predict based on errors, here, the AI-based system was trained using fine granular data that are no longer available later, thus resulting in poorer performance. This provides multiple challenges along the FATE characteristics. The poor performance can lead to suboptimal recommendations, and depending on the level of transparency provided initially, questions related to accountability between the system developer and system provider can arise. Creating transparency regarding the training data's origins and the data used for the recommendations helps mitigate this issue.

Data repurposing. In addition to biases resulting from sampling, data collection practices also introduce misuse and biases. Traditional data collection practices differ significantly from contemporary practices in AI systems development [33]. The traditional practice is to collect data for a specific purpose. For example, a clinical trial of a drug used to treat COVID-19 will collect experimental data to assess the drug's side effects.

However, repurposing data is the norm in AI-based systems. For example, a blood test result in a patient's electronic healthcare record that has been captured to diagnose a certain disease may also be used by an AI-based system to diagnose other diseases. This can be a potential issue compromising the accountability characteristic of the algorithm. For example, while the quality of data from medical images can be sufficient for the original purpose, such as stroke detection, it may not meet the needs of subsequent data uses, such as finding new disease markers [5]. Repurposing data creates ambiguity about the data and their origins, making it difficult to clearly identify the person or entity accountable for any incorrect recommendations.

Data augmentation. When the available dataset is not large enough for the intended computations, data augmentation might be used (i.e., increasing the size of the dataset with synthetically generated data or slightly modified copies of the existing data, for example, through translation, rotation, flip, or scale). For instance, augmented data are generated through the rotation, translation, and scaling of a prior dataset on liver lesions [26] when training a generative adversarial network (GAN). These modifications and the synthetically generated data can amplify existing biases within the dataset and mask the inadequacies of the collected data.

Some AI algorithms rely solely on simulated data. For example, AI systems have been developed to design bridges and control robot arms using only simulation data [23]. Simulations can create useful data to learn from, especially when little input and manually labeled data are available. However, because deep learning can approach problems more intuitively by focusing on patterns in the core data, researchers have suggested that AI systems perform better without synthetic additions to the data [23].

Therefore, data augmentation and the use of simulation data bring about new challenges to the fairness and accountability characteristics of AI algorithms. Data augmentation amplifies existing biases and creates opaqueness about the actual representativeness of the data, thus limiting transparency and making it more challenging to identify the cause of an incorrect recommendation.

2.2 Biases from Data Pre-Processing

Data processing is vulnerable to errors that introduce biases, such as dataset shifts, opaque pre-processing, data labeling, adversarial manipulation, transfer learning, and data augmentation.

Dataset shifts. An easily ignorable fact is the non-stationary nature of the environment and the population from which all the input data of AI-based systems are generated [59]. For example, when a data shift occurs, an important predictor of a specific disease at one point in time can be more or less important at a later point in time because of improvements in the quality of care available. For instance, many predictions using the Medical Information Mart for Intensive Care dataset are confounded by changes in hospital operation practices [71]. Considering time as an influential variable shows dataset shifts caused by changing practices, which, in turn, result in significant changes in the observed data. Unless this data shift is identified and the AI algorithm is retrained or recalibrated [53], the performance of the system deteriorates, affecting the fairness, transparency, and explainability characteristics of the algorithm. Low performance can lead to incorrect recommendations that negatively affect users. If the data's origins and subsequent changes in the environment are not made transparent, the derived explanations will be at least distorted.

Opaque pre-processing. AI-based systems are often characterized as black boxes [2]. While some AI-based systems provide accurate predictions, the rationale behind their predictions remains opaque. In algorithms with intrinsic obscurity, such as deep neural networks, understanding the specific patterns being learned is difficult [53]. For example, in a study detecting hip fractures, an algorithm was confounded by the scanner model and by scans marked as "urgent" [8]. Therefore, assessing the potential biases introduced when using the output of an opaque algorithm as an input for another AI-based system is difficult. Opaque pre-processing limits the transparency and explainability of AI-based system recommendations. If it is unclear what data were used to train the system, confounding indicators are more difficult to identify and assess, and they do not allow users to learn relevant insights. However, deriving explanations for the recommendations can help experts validate the model and its recommendations. Different types of explanations (e.g., feature extraction, pre-defined models, and sensitivity [87]) can help an expert evaluate, improve, and correct the model.

Data labeling. While data quality chasm refers to data that may appear to be similar but have different qualities, another issue arises with data labeling, as the identification and development of labels are often not transparent. Data labeling is related to supervised learning, such as medical image classification. The outcome labels are used by supervised algorithms in the training stage. While automated techniques for data labeling (e.g., with weak supervision) are on the rise [50,76], labeling is often a labor-intensive task and is frequently performed by unqualified or poorly trained ghost workers or through crowd-based platforms [7].

Incorrect labels in the training data create erroneous or unfair recommendations and explanations developed by AI-based systems because of the inherent bias embedded in the training data. This bias affects the fairness, transparency, and explainability characteristics of the AI algorithm. Fairness is affected, as unqualified or poorly trained ghost workers will make mistakes and possibly bring their social biases into the data. As these are undesirable business practices, organizations seldom disclose them, thereby negatively affecting transparency. While these business practices introduce biases, hiding them from customers makes it difficult for both the user and the expert to benefit from explanations.

As the majority of existing data are non-labeled and are usually very expensive to label, some researchers perceive the reliance on labeled data as even counterproductive to the development of effective AI [23]. A recent trend in the automatic labeling of data using AI [77] has emerged. The idea is simple. As labeling is often a bottleneck task in AI system development, we could use machine learning (ML) to extrapolate the labels. A labeling ML algorithm can be trained based on a limited number of available or easily attainable labels and can then be used to label a larger dataset. While this reduces the effort of manual labor, it may also increase the severity of biases already existing in the smaller sample, leading to erroneous or unfair recommendations and explanations.

Adversarial manipulation. As AI-based systems derive their models based on nuanced variations in the data, sometimes, small changes in the data input can lead to significant differences in the output [38]. Therefore, AI-based systems are potentially susceptible to adversarial manipulation. For instance, images of benign moles may be misdiagnosed as malignant because of added adversarial noise or seemingly minor changes in the data [53]. These manipulations can be intentional, such as when an attacker changes the input of an algorithm to fool it, or unintentional, such as when a user accidentally rotates an image used as an input. Without sufficient transparency of the data preprocessing, identifying this potential threat in an otherwise effective model is difficult. These seemingly minor changes can result in significantly different outcomes that make explaining the recommendations difficult and the recommendation itself possibly incorrect.

Transfer learning. Once an AI-based system is built, we may use the algorithm to solve similar problems. In particular, a new AI-based system benefits from the information learned from another system. For example, a pre-trained model can be used to encode radiographic features in images before final re-training [8] to improve the sample efficiency for a reinforcement learning agent. Transfer learning can also improve AI system performance when predicting cancer for ethnic groups with limited data availability [29]. However, transfer learning only works when the source task is closely related to the new task. If not, transfer learning introduces biases and negatively affects performance [91]. As transfer learning also increases ambiguity about the AI-based system's recommendations, it impedes clear accountability. Therefore, transfer learning should be made transparent to the user, as it otherwise adds to the system's opaqueness.

3 Recommendations for Implementing Data Provenance

Considering the importance of mitigating data-induced biases originating from data sources and data pre-processing, organizations need to establish data provenance when implementing responsible AI-based systems that address the FATE characteristics. We propose a data provenance framework for responsible AI to enhance its FATE characteristics (Figure 1). Organizations can focus on three key areas: establishing organizational data governance, demanding data traceability, and leveraging technological advances, such as explainable AI. Below, we summarize current and future challenges and elaborate on actionable recommendations and how these enhance the specific characteristics of responsible AI (see Table 3).

Table 3 – Overview of the current state, challenges, and recommendations

Current state	Challenges	Recommendations
Organizational data lineage and accountability are lacking.	Governmental organizations demand control and protection of data integrity, confidentiality, and availability.	**Establishing Organizational Data Governance:** - Managing meta-data - Conducting data audits
Organizations rely on data from multiple data sources in their AI systems, creating heterogeneity and opaqueness. **Many current AI-based systems rely heavily on manually labeled data.**	Organizations typically do not have a clear understanding of the source and processing of data, such as various experiences, goals, and perspectives of the people annotating the data.	**Demanding Data Traceability:** - Guiding data acquisition - Benefitting from blockchain technology
Technologies seek to increase the transparency of AI models.	Little attention has been given to data opaqueness.	**Leveraging Technological Advances for Data Provenance:** - Deriving rules for explanations - Identifying possible adversarial manipulations - Finding the inherent structure in the data

3.1 Establishing Organizational Data Governance

Several governmental organizations have launched directives, laws, and regulations to provide control and protection of data integrity, confidentiality, and availability. Examples include the US Health Insurance Portability and Accountability Act (HIPAA) and the EU's General Data Protection Regulation (GDPR). However, current data governance practices are often limited to master data management, that is, a set of processes related to the who, what, and where of business transactions, communications, and events. Seemingly, organizations too often mimic what their competitors do rather than being proactive and shaping the course of action. For example, many organizations are still seeking to become data driven. Yet, once they achieve this, they find that inadequate attention is given to data governance during the development of AI systems, which, in turn, creates additional challenges [44].

Organizations need to establish organizational data governance practices that enforce data lineage and accountability. This would help them not only meet increasingly strict regulatory requirements but also benefit from an overarching perspective of their data assets. Particularly, organizations need to manage their meta-data and conduct data audits in order to respond to the organizational challenges associated with inadequate data governance.

For some organizations, these goals stand in a potential conflict. For example, data privacy seeks to protect individuals from being identified—often through personal identifiable information—or being associated with such information. Data lineage, on the other hand, refers to the visibility of the data's origins and further processing. If the data's origins and further processing are done by individuals, both concepts stand in conflict. An organization will have to manage this conflict by enhancing responsible AI under the condition of privacy policy compliance, such as the GDPR [98]. For example, an organization may allow identifiable data to be traced only for specific legal purposes. Organizations also need to leverage some privacy-preserving approaches, such as federated learning, to allow the safe sharing of identifiable data or models across entities [69].

Managing meta-data. Meta-data describe data and consist of detailed information about the data captured in a data source. Meta-data help maintain the data within an organization in a manner that ensures the timely, efficient, and accurate retrieval of the required information [68]. It also helps ensure that processes and activities are documented in a transparent and verifiable way [78]. Generally, there are two practices that organizations use to manage meta-data: cataloging data and curating data. A data catalog stores information about the data, such as the rationale for choosing a data source, the stakeholders involved, and the content stored within it. Such information may also be documented in a datasheet [31].

Extending these efforts, organizations should establish clear processes and responsibilities for data curation. Data curation identifies and leverages the data within the organization and helps assess the FATE of system recommendations. For example, organizations can identify representation and corresponding limitations by visualizing and clustering data annotations. These annotations facilitate the identification of discriminatory correlations between features, labels, and groups.

Overall, managing meta-data through data catalogs and data curation helps increase the benefits of existing data through increased transparency [68] and helps reduce costs by avoiding unnecessary data collection. Managing meta-data also requires clear accountability for the different data sources. Meta-data help organizations benefit from transformation, weighting, and sampling techniques [4] by minimizing the extent to which data deviate from the objectives of responsible AI, thus helping ensure fairness of the recommendations.

Conducting data audits. Enhancing data auditing capability in an organization is another approach to establishing data provenance through data governance [44]. Data auditing is the process of assessing whether the data are fit for a specific purpose. Given the recent increase in regulatory requirements, organizations should conduct data audits to assess the data used within their systems, similar to the way they assess and audit other aspects of their business operations. Data audits help uncover potential biases related to data processing and their associated consequences. With a reasonable and suitable guarantee of authenticity and reliability, data audits help enhance the accountability and fairness of AI-based systems. This not only applies to high-reliability organizations that need to make high-stakes decisions but also provides benefits for other organizations that seek to act responsibly. Data audits consist of data profiling (e.g., assessing the availability and quality of data and the risks associated with data integration [45]) and impact analysis (assessing the impact of poor data quality on performance and profits) [57].

Data audits become increasingly important when individual services are integrated into larger systems [73]. Conducting data audits enhances the fairness of AI systems by ensuring a good fit between the data and their use. Conducting data audits also requires clear accountabilities for the appropriate handling of data. In addition to establishing data accuracy, data audits uncover data silos and areas where more depth and/or breadth of data is necessary for the AI-based system to provide valid recommendations. A data provenance record could document the data capturing and data processing entities for the dataset in question, simplifying the audit process. Data provenance records also help in understanding the data's origins and pre-processing, thereby enhancing transparency.

3.2 Demanding Data Traceability

Managers need to be aware of the implications of using different data sources and processing methods, especially when they seek to achieve fair and transparent systems. Data traceability is gaining increasing attention as managers become aware of its importance. For example, it usually takes Walmart 6 days and 14 hours to identify the source of a farm product. When the supply chain data are maintained in a blockchain, however, it takes only 2.2 seconds to establish complete data traceability. Therefore, platform providers need to enhance the traceability characteristic of data provenance in order to improve the efficiency of business and decision making.

Enhanced traceability provides more information about the historicity of data and increases overall transparency. Transparency enables the creation of an intermediate representation of the original data [4] encoding the responsible AI objectives, such as fairness. As a result, organizations mitigate biases resulting from data sources and improve the fairness of their systems. Demanding data traceability may include guiding data acquisition and leveraging blockchain technology.

Guiding data acquisition. Many current AI-based systems rely on manually labeled data. Despite the recent trend of increasingly using automated labeling practices, manual labeling is still indispensable. Manual labeling either applies to the entire dataset or only a subset of datapoints for later extrapolation. Either way, if organizations do not have a clear understanding of the various experiences, goals, and perspectives of the people annotating the data, they cannot account for the significant impact on data quality [51]. Organizations should develop procurement guidelines that take the traceability of data into consideration. For example, managers need to demand transparency regarding data origin and quality when acquiring external training datasets. A data provenance record identifies the true source and subsequent processing of data, uncovering the often-hidden history of the data. Recent end-to-end provenance projects have developed a set of tools, such as R packages, that allow organizations to establish data provenance through enhanced data traceability [24].

Furthermore, some data used to train the system may not have been labeled by experts, whereas other data may have been procured from data brokers (organizations that collect data for the purpose of reselling them). Understanding the sources and methods used to acquire the data is critical to ensure that they are ethically and legally collected (e.g., with informed consent). Demanding traceability (e.g., through a data provenance record) increases transparency and helps organizations identify the accountable actors for mitigating risks related to the use of AI-based systems' recommendations.

For instance, an organization should provide the descriptive statistics of a dataset as part of its data provenance records, allowing users to identify the potential risk for discrimination. Based on these statistics, users can evaluate the AI-based system's recommendations to correct, mitigate, and avoid future discrimination, either by altering the input data, modifying the algorithm, or changing the way in which predictions are made [4]. As a result, the user is likelier to perceive the recommendations of the AI-based system as fair.

As data provenance relates to a record of the data's origins and subsequent processing [9], it also increases transparency. For example, data provenance is needed to develop a data information sheet [31] that provides details on the most important variables influencing an AI-based system's recommendations. As such, data provenance provides users with basic information about the data and their processing [17] before they are used by the AI black box. A user can ascertain whether the data used to train the system are suitable and relevant [36].

Benefitting from blockchain technology. Blockchain-based data provenance is a promising approach to enhance the traceability of data in responsible AI. Blockchains can record the meta-data and history of data objects. The important characteristics of blockchains, such as transparency and auditability, enable the security and traceability of the meta-data, which are crucial for data accountability. Data immutability in a blockchain also enhances the perceived fairness in the recommendations. Various data provenance architectures based on blockchain technology, such as ProvChain [60] and LineageChain [60], have been proposed. Blockchain technology has also been leveraged to handle dark data [99], which are the data that organizations collect but fail to utilize for their value. As a secured distributed ledger, blockchain has the potential to upgrade the value of the data and provide more efficient and transparent results [70].

Increased transparency supports a consumer-centric strategy that organizations increasingly follow. For example, in healthcare, the notion of patient-centered care refers to being respectful and responsive to individual patient needs, values, and preferences; this requires health IT systems to prioritize data provenance and the transparency of patients' personal health-related data. With increased transparency, patients are better informed and are therefore more empowered to seek clarification on diagnoses or recommendations [41]. This interaction improves the quality of healthcare. It also enhances patients' confidence in the care provided and hence its effectiveness. Healthcare organizations' attention to data provenance in electronic healthcare records improves the transparency of their decisions and recommendations.

3.3 Leveraging Technological Advances for Data Provenance

Given the opaque nature of many AI-based systems, data provenance is essential for understanding AI-based systems' recommendations [74]. Recent technological advances include explainable artificial intelligence (XAI) methods, GANs, and deep learning with advances in small data techniques.

Deriving explanations. XAI methods, such as LIME, LORE, and Anchor [29], push the traditional boundaries imposed by trade-offs between the accuracy and interpretability of AI systems' recommendations. More recently, XAI solutions have allowed users to understand the most important features that lead to the outcomes, make changes to model features, and customize the model explanation [58].

Explainable AI methods seek to increase the transparency of AI models, but little attention has been given to addressing data opaqueness. Data provenance provides a complementary perspective toward transparency for the user [6] by presenting information about the source and further processing of the data used to feed an AI-based system. Data provenance helps provide complementary information to the explanations provided by XAI systems. For example, expanding the data provenance concept to AI algorithms facilitates the documentation of the data processing performed by an AI algorithm through global and local explanations [22]. While a global explanation creates transparency regarding the model used to make all recommendations (e.g., answering the question of how the AI makes its recommendations for all patients), a local explanation provides transparency for a specific recommendation (e.g., answering the question of why the AI makes a specific recommendation for a particular patient). For example, through explainable AI, healthcare providers and their patients can better understand the important factors that lead to an algorithm's recommendations on a particular diagnosis or treatment, thereby enhancing the accountability of the parties responsible for and receiving care. Therefore, we suggest that organizations should strive to make the most of recent technological advances related to XAI.

In particular, we suggest that organizations should leverage existing XAI methods, such as LIME and LORE, and XAI techniques, such as layer-wise relevance propagation [85] and gradient-based explanations, with supporting architectural frameworks, such as CaSE [55], to provide easily understandable explanations of AI-based recommendations. XAI methods, for example, derive rules that explain how a recommendation was reached by presenting cut-off values that lead to the predicted outcome or by identifying the factors that most strongly influence the recommendation. Such explanations help users better understand the AI system's behavior and identify new patterns in the data.

However, prior studies also suggest a potential conflict between explainability and other FATE dimensions. For example, a trade-off exists between explainability and fairness [56]. While explainability seeks to simplify the complex nature of AI-based systems so that they can be understood by humans, there is an inherent loss associated with this simplification that may lead to new biases. Organizations can manage these conflicts, for example, by using multi-criteria decision-making methods (see [89] for an overview) to guide and prioritize different characteristics. In a given scenario, one characteristic might be more important than another. For example, if the adoption and use of the system are concerns, explainability could be one way of increasing the transparency of a system to increase trust [80]. In organizations that provide a process for users to participate in the evolution of the system in order to address potential fairness concerns [42], users are less likely to reject the system.

The lack of explainability of AI prediction outcomes can be caused not only by the black box nature of algorithms but also by the biases in the data. While most research focuses on algorithm explainability, we suggest paying additional attention to how data provenance can enhance the explainability of outcomes. By allowing individuals to meaningfully interact with the system and by enhancing the explainability of AI-based systems, organizations facilitate autonomous decision making, detect errors, minimize biases, and thus safeguard justice [15].

Managing noisy data. The presence of meaningless and irrelevant data is often referred to as noise within the data. Scholars have made significant progress in managing noisy data that organizations can benefit from. A distinction is made as to whether the noise relates to predictive attributes (referred to as attributed noise) or to target attributes (referred to as class noise). Different techniques are available for identifying and handling noise within the data. A recent systematic review provides a good overview of the current state of the art on the problems caused by noisy data in AI-based systems [40].

The management of noisy data is important for deriving fair recommendations. In fact, striving to achieve fairness without addressing the noise within a given dataset could backfire. For example, a prior study investigated the use of noise models for denoising data during subset selection [65]. Scholars applied noise models to select a subset of data from an existing larger data set. The goal was to generate a fair dataset so that the sub-dataset accounts for race while having noisy race data. The study points out that failing to account for noise has unintended side effects, as it decreases the fairness of the resulting subset selection.

Different techniques are available to handle noise within data [40]. For example, organizations can use filtering techniques to identify and remove noise, or they can alter the data, sometimes referred to as data polishing. They key difference between responding to class noise and to attribute noise is that for class noise, organizations should also consider relabeling, whereas for attribute noise, organizations can use data imputation.

A related technique is the use of GANs (sets of neural networks that seek to generate new data with similar characteristics as the training data). Organizations should use GANs to identify possible adversarial manipulations, thereby mitigating negative consequences. For example, GANs are used in image-to-image translations, such as the translation of low-dose CT scans that have noise in the data into regular-dose CT scans. In this case, a generator network translates the low-dose scan into a regular-dose scan, whereas a discriminator tries to distinguish the artificial from real regular-dose scans. As a result, the noise in image-to-image translation is reduced [96].

Identifying inherent data structures. Deep learning for text, audio, and video recognition often involves performing a pre-text task to find an inherent structure in the data of their AI systems. The pre-text task is self-supervised learning with the purpose of generating a useful feature representation for the downstream task [12]. Pre-text tasks may force ML models to deconstruct data in order to enhance explainability [23]. For example, the Facebook AI Research group uses a combination of clustering and training based on rotated images to improve the quantity of unlabeled data used in their image classifier. After this pre-text task processing, the second stage of training uses conventional labeled data to create interpretable results [23].

Furthermore, advances in small data techniques help organizations improve the performance of AI-based systems. While many AI-based systems rely on large data, some of the most valuable datasets are only available in small quantities [51]. For example, the application of AI in the medical domain often requires data labeling by medical professionals, such as radiologists or physicians. A review by a radiologist is needed to reliably label an image scan with the correct diagnosis of the presence or absence of lung cancer. As medical professionals' time is scarce and expensive, and the task of data labeling is quite repetitive, the creation of large datasets is a challenge. However, it is this high-quality human input that facilitates high-quality recommendations by AI-based systems.

Overall, a clearer understanding of the system's behavior and the data helps judge the fairness of recommendations. This is important because, for example, evidence-based medicine rests on high standards of explainability, as medical decision making requires a sound understanding of underlying disease mechanisms and treatments under particular conditions [88]. The lack of this understanding undermines the implementation of AI in healthcare [81]. This issue is crucial because of the promising benefits provided by AI in healthcare.

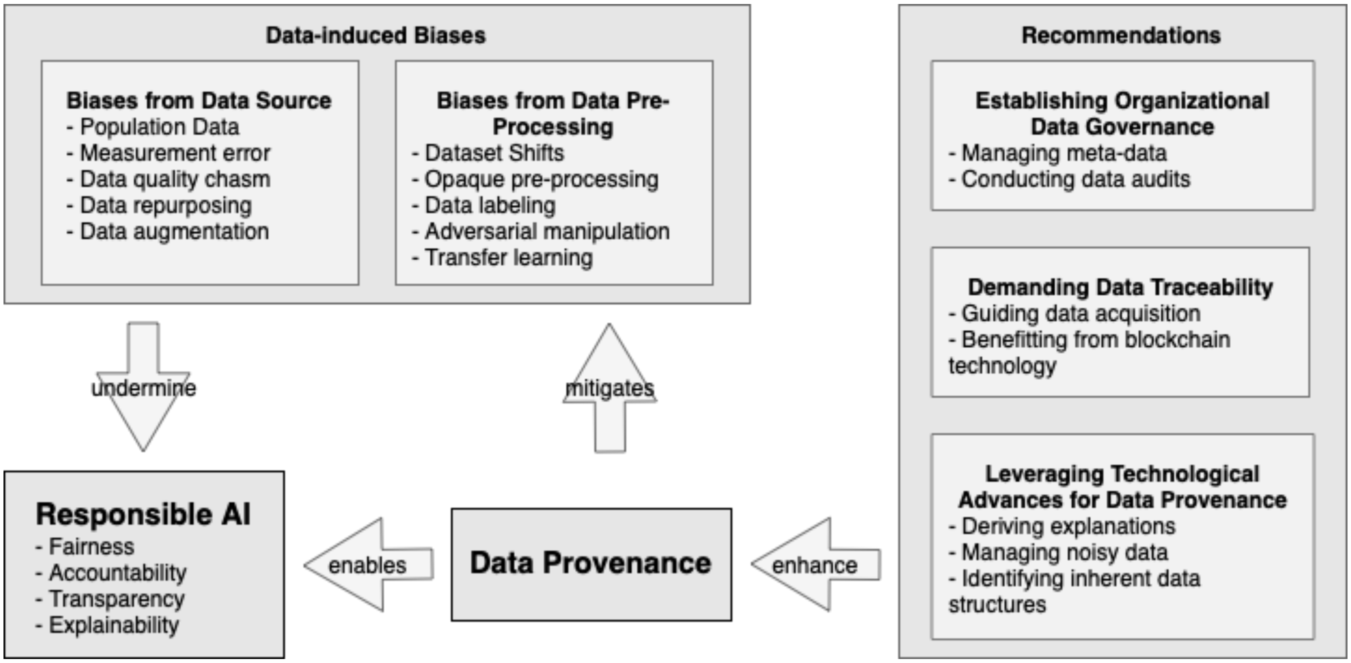

Figure 1 – Data provenance framework for responsible artificial intelligence

4 Exemplar Application of the Data Provenance Framework

We discuss the application of our framework with two recent examples that highlight the problems associated with a lack of responsible AI.

A recent example of data provenance concerns relates to the application of AI recommendations in healthcare. A recent study evaluated the performance of the **AI-based system that is embedded within EPIC** [100], a major electronic healthcare records system, to predict sepsis (a potentially life-threatening condition in which the body's response to an infection damages its own tissues). As sepsis is the number one killer in US hospitals [67], hospitals attach great importance to identifying and treating conditions that may lead to sepsis. There is widespread adoption of sepsis prediction models, such as the one provided by EPIC. However, the study suggests that i) the AI-based system does not deliver the advertised performance, ii) important assumptions that underly the AI system require careful examination, and iii) the system's high number of false positives contribute to alert fatigue for the medical staff [100].

This case highlights four important biases: data repurposing, population bias, transfer learning, and data shifts. One important observation of the evaluation was that the data used in the development of the model may have been repurposed. To derive the predictions, EPIC measured positive sepsis cases based on billing codes but not on the clinical definition of sepsis. The decision to use billing codes also results in population bias, as the presence of sepsis relies on the identification of sepsis by the medical staff. Yet, the medical staff used the system with the expectation that it would help predict sepsis before medical personnel could identify it. In response to the study, EPIC has argued that transfer learning could explain the suboptimal performance. That is, transfer learning works only when the source task is closely related to the new task, so the sepsis prediction model developed using the data from one environment may not work well in other environments. Transfer learning may have introduced biases, negatively affecting the performance of the sepsis prediction model using data from the University of Michigan Hospital [94] in contrast to data from the University of Colorado Hospital [10]. Lastly, the researchers also describe the potential for a dataset shift resulting from changed practices in treating sepsis and suggest the need to retire old models entirely. We suggest that organizations using prediction models such as this should establish organizational governance, conduct data audits, and leverage technological advances in the area of XAI to derive explanations for these prediction models.

Organizational data auditing capability establishes data provenance through data governance [44], whereas data auditing refers to the process that assesses the fit of the data for a specific purpose. A data audit would allow healthcare organizations to evaluate the data used to train the AI system and identify possible concerns. In our example, a data audit would allow a medical expert to identify potential errors resulting from the use of billing codes as a proxy for the presence of a disease. Yet, billing codes are used in the administrative process and can deviate from the medical diagnosis (e.g., [93]). When used in research, billing codes are often a means to identify patients for another study in order to narrow down those who are likely to have a specific disease or condition (e.g., [86]).

An organization's capability to audit AI systems has become increasingly more pressing, as a recent study suggests a severe lack of transparency by AI system providers and a lack of oversight by the FDA [67]. Medical experts criticized the opaqueness and limited transparency offered by EPIC. As the AI system is protected by intellectual property rights, the developer has disclosed very limited information about the development of the prediction model. Medical professionals implicitly relied on the FDA's oversight, but the recent study points out that the FDA's oversight is limited [101]. Medical devices are rated by the FDA into three classes [16], with the highest class being reserved for life support systems. Those systems that make autonomous decisions (e.g., a pacemaker or an automated insulin pump) are required to meet the highest standards set by the FDA. AI-based systems that provide recommendations to healthcare providers (e.g., a sepsis prediction model) are often considered class II systems that have much lower FDA oversight. In the EPIC example, the study suggests that not even the reduced oversight was applied, as the system may have been checked upon market launch, but later additions are not subject to further FDA approval.

Recent technological advances help organizations identify the needed adjustments. For example, explainable AI helps provide insights and feedback to AI developers so that they can then further refine the AI system by adjusting the network architecture or retraining the model. This concept is often referred to as *human-in-the-loop* and has been advocated by scholars for the debiasing of AI systems [47]. Here, the technological advances in XAI can enhance data provenance by supporting feedback through human-in-the-loop and, in turn, improve the transparency of the predictive model. For example, a medical expert could question the validity of the model for the early prediction of sepsis, while the most important prediction factor of the trained model is, in fact, the diagnosis of sepsis by medical staff (i.e., labels of the training data).

The concerns described are not limited to the healthcare domain. Another example is the **Amazon AI recruitment tool,** which has received attention for its lack of adherence to the facets of responsible AI (e.g., [46]). Amazon developed an experimental hiring system that was designed to automatically screen the resumes of job applicants and identify the top candidates. Amazon later realized that the AI system did not select candidates for technical jobs in a gender-neutral way but was rather biased negatively toward female candidates. In hindsight, the explanation for this behavior seems obvious. It was reported that among Amazon's entry- and mid-level corporate employees, women accounted for 31% of the workforce last year [52]. The system had been trained with data for the past 10 years, during which male candidates were predominantly chosen for technical jobs. Meanwhile, many high-technology companies have realized the gender discrepancy when hiring employees for tech jobs and have changed their hiring practices to recruit more women. In such cases, the data shift would require the developers of AI systems to discard older data and rely on more recent data to train their models.

Amazon used its own recruitment data from the past 10 years in training the system. An auditing process would have helped enhance data provenance and thus uncover the presence of a dataset shift and population bias. Specifically, it would have highlighted that the hiring practices followed during the past 10 years have been significantly unfair to female candidates [54]. Further adjustments are necessary to ensure responsible AI recommendations. Thus, data auditing can help increase the fairness of a system by establishing data provenance.

In a similar vein, the human-in-the-loop that has been advocated for debiasing HR recruitments systems [47] helps organizations evaluate the AI system. Technological advances in XAI enhance data provenance by supporting feedback through human-in-the-loop and, in turn, help mitigate the negative impact of a dataset shift. XAI enhances the explainability of responsible AI through data provenance.

5 Research Agenda

Organizations continue adopting and using AI-based systems to support evidence-based decision making. A particular focus is on enhancing the FATE of the implemented AI-based systems. Our review of data-induced biases and discussion of how organizations can mitigate these by establishing data provenance within their organizations lead to three central recommendations for organizations. Yet, more research is needed to improve data provenance methods, tools, and practices for responsible AI. Thus, we develop recommendations for future research, identifying four central topics (see Table 4).

Table 1 – Exemplar research questions for responsible AI

Research topic	Exemplar future research question
Conceptual Clarity	How can we classify central terms related to data provenance and responsible AI? What are the relationships between AI explainability and AI interpretability? What are the relationships among FATE and what are the boundary conditions for the impact of date provenance on the FATE of responsible AI?
Resolving Tradeoffs	What are the existing tradeoffs or conflicts among the goals of responsible AI, and how can we resolve them? How do different organizational profiles affect the design of responsible AI in organizations?
AI ethics	What are the regional differences in moral and legal concerns that impact responsible AI? How do we ensure responsible AI with increasing role of AI in the future of work? How do we develop and implement scalable, responsible AI solutions?
Designing responsible AI	What are the design guidelines and principles for responsible AI systems? How do we design explainability to enhance interpretability, and what are the influential conditions?

Conceptual clarity. Establishing a clear nomological network to better understand the distinction of terms and their relations is crucial for the development of data provenance for responsible AI. More research is needed to determine the unique nature of different concepts and possibly the interchangeability of some concepts. Scholars can use taxonomy development methods to identify classifications with mutually exclusive and collectively exhaustive dimensions. For example, explainability and interpretability are essentially two related but different concepts but often are used interchangeably; terms such as data lineage and data pedigree are closely related to data provenance, but they are distinct terms. With enhanced conceptual clarity, more research can be conducted to understand the relationships between ontologically different concepts.

Understanding the conditions in which these relationships occur is also important. For example, having a fair dataset or fair recommendations does not necessarily guarantee high transparency. This can help explain conflicting evidence in existing research. For example, regarding the relationship between transparency and explainability, some scholars suggest that explainability enhances the transparency of systems [82], whereas others argue that explainability is a sub-characteristic of transparency [83]. More research is needed to develop a nomological network around data provenance for responsible AI.

Resolving tradeoffs. Implementing data provenance for responsible AI can lead to tradeoffs or conflicts. For example, regulations, such as the GDPR, require the system to ensure data privacy, whereas other requirements demand more traceability, such as auditing requirements. The case of Twitter's cropping algorithm shows a conflict in speed and consistency versus the risk of making incorrect predictions [95]. Furthermore, the trade-off between accuracy and interpretability is an often-mentioned conflict related to responsible AI [90]. More research is needed to identify these conflicts and develop corresponding resolutions. Researchers can benefit, for example, from specific research methods, such as conjoint analysis [37] and analytic hierarchy process (AHP) approach [79], in order to prioritize different characteristics or identify important configurations of characteristics in different context.

In order to resolve these conflicts, we suggest two important avenues. First, scholars may benefit from research on multi-criteria decision making. Prior research can guide managers in making decision while accounting for multiple and potentially conflicting goals. These require extension and evaluation for responsible AI before they can be used to derive normative recommendations. Second, organizational or AI project profiles may be created to provide templates for developing responsible AI projects. While prioritization may be the result of external forces, such as governmental regulations, they may also be the result of organizational values and culture. For example, an open and progressive organization may prioritize transparency and fairness over accountability concerns. In contrast, a risk-adverse organization may focus on accountability and performance over transparency. Similarly, different projects within an organization may need to emphasize different aspects of FATE. Future research could explore the role of organizational and AI project specific profiles in the development and use of responsible AI systems.

AI ethics. Questions related to the fairness of responsible AI are often at the cross-section of research focused on novel technology and its ethical behavior [64]. Research related to ethics is closely associated with moral and legal questions. Legal research is often conducted at the national level according to the local needs of the judiciary system. By contrast, new technical challenges emerge during the development and deployment of responsible AI-based systems regardless of local needs. For example, responsible AI has the potential for solutions that are easily scalable from a technical perspective yet raise concerns when it comes to local legal requirements, such as the GDPR.

Prior research also coined the term *responsibility gap* [49], describing a situation in which artificial agents are used to decide on a course of actions or in which they act themselves without human involvement. As the rules by which they act are not inscribed during development, there is no individual who assumes responsibility for the machine's actions. Current ethical and legal frameworks have not been designed for these situations, leading to a responsibility gap [63]. In addition to mitigating or eliminating the responsibility gap, organizations must often follow multiple goals, such as transparency and accountability [21], in the development of responsible AI systems.

However, how governmental regulations that organizations must follow map toward different goals of responsible AI remains unclear. For example, future research should investigate whether and how we need to extend and modify regulations, such as HIPAA in the US and the GDPR in the EU, to allow platform providers to offer scalable yet responsible AI solutions.

Designing responsible AI. Designing responsible AI provides a particular challenge for future research, as it requires us to instill human and social values into the AI system in a way that users see and appreciate it [21]. However, current research often focuses on the technical implementations of FATE. For example, much research related to explainable AI offers technical solutions for developing explanations. When an explanation is presented to the user, an interpretative process is triggered. The user will develop an autonomous interpretation of the explanation, a process that is often described as the interpretability of an explanation. This interpretation may or may not be in line with the expected interpretation intended by the system's designer.

Therefore, more research is needed to better understand the link between different design patterns and technological solutions related to explainability research and the interpretability of individual users. For example, certain user or task characteristics influence the interpretability of a user in the sense that an expert, compared with a novice, requires different explanations. We suggest that data provenance requires also more attentions, particularly in the XAI community, as it provides important complementary information that are crucial for the interpretation by the user. Future research could develop clear guidelines, design features, and design principles for designing responsible AI systems,

6 Conclusion

Data provenance is important to mitigate biases and improve responsible AI-based systems (see Figure 1). Existing practices view data provenance as a mandate of directives, laws, and regulations designed to ensure the control and protection of data integrity, confidentiality, and availability. Data provenance is viewed as the cost of staying compliant with these requirements. Such practices result from a lack of organizational commitment to developing responsible AI-based systems.

By contrast, our recommended practices view data provenance as an important component of developing responsible AI-based systems. Organizations that are strategically committed to their FATE goals are likely to achieve long-term improvements in organizational performance. Our recommended practices view data provenance as an investment necessary to meet their FATE goals and recognize that the loss of data provenance at any point in the provenance chain leads to a loss of data provenance in all subsequent parts. Therefore, organizations need to recognize the importance of establishing a comprehensive provenance for critical data that serve as inputs to AI systems.

In contemporary systems development projects, such as in the case of data-driven development and AI engineering, data repurposing is becoming more and more the norm. Recommended practices will help organizations benefit significantly from data provenance, as the data provenance established for one project is likely to benefit several other projects that use the same data. Therefore, when examining the costs and benefits of data provenance, organizations need to take a comprehensive view that spans across projects, as different projects often draw from the same data sources. Whereas existing practices view data provenance records as static, recommended practices recognize the need to maintain dynamic data provenance information that is updated throughout the data's lifecycle.

We have outlined the multiple benefits of data provenance along and beyond the FATE characteristics. However, organizations will need to prioritize their investments in data provenance efforts based, for instance, on the magnitude of benefits resulting from achieving FATE and the severity of negative consequences or the cost of failure that result from not achieving FATE. Organizations that view data provenance as an overhead cost are likely to neglect it when operating under budget or schedule constraints and, even worse, perhaps engage in undesirable practices, such as virtue washing [92].

Investments in data provenance should be driven by an intrinsic motivation to improve the responsibility of AI-based systems. For example, adopting data provenance practices to achieve transparency is valuable because it enables users to understand, engage with, and audit the AI-based system and its outcomes. Similarly, data provenance that enables accountability is a means to ensure justice by clarifying responsibility and avoiding harm from deterrence [15]. As these examples show, FATE characteristics are instrumental in upholding the intrinsic values of core principles, such as human autonomy and justice. In addition, organizations that take a lifecycle perspective recognize that the costs incurred in the early phases of data acquisition and processing lead to benefits later in the AI-based system lifecycle. Yet, these benefits, such as increasing reputation, avoiding the loss of reputation, and establishing the desired FATE characteristics, are often difficult to quantify despite quickly outweighing negative implications.

In high-reliability organizations, such as healthcare providers, suboptimal decisions can have severe consequences. The increasing reliance on AI-based systems and the lack of understanding of the data used to generate recommendations highlight the importance of data provenance. Establishing data provenance guidelines and policies can facilitate the FATE of AI-based recommendations. For example, in the context of the

COVID-19 pandemic, the provenance of data is important for discerning the FATE of recommendations made by AI-based systems that rely on data from varied and disparate data sources. While more guidelines are needed to develop data provenance throughout the entire data lifecycle [11], implementing the recommended practices is an urgent task for organizations that aim to harness the benefits of AI-based systems. Our recommendations will help organizations enhance essential data provenance capabilities toward fair, transparent, accountable, and explainable evidence-based decision making by responsible AI-based systems. Our proposed research agenda suggests potential research avenues related to data provenance. We suggest that achieving conceptual clarity, resolving tradeoffs, observing AI ethics, and designing responsible AI require more research by scholars from different disciplines.

REFERENCES

[1] ACM U.S. Technology Policy Committee. 2020. Statement on Principles and Prerequisites for the Development, Evaluation and Use of Unbiased Facial Recognition Technologies. Retrieved August 24, 2021 from https://www.acm.org/binaries/content/assets/public-policy/ustpc-facial-recognition-tech-statement.pdf

[2] Amina Adadi and Mohammed Berrada. 2018. Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI). IEEE Access 6, (2018), 52138–52160. DOI:https://doi.org/10.1109/ACCESS.2018.2870052

[3] Gediminas Adomavicius, Jesse Bockstedt, Shawn Curley, and Jingjng Zhang. 2019. Reducing Recommender Systems Biases: An Investigation of Rating Display Designs. MIS Quarterly 43, 4 (February 2019), 18–19.

[4] Gediminas Adomavicius and Mochen Yang. 2019. Integrating Behavioral, Economic, and Technical Insights to Address Algorithmic Bias: Challenges and Opportunities for IS Research. SSRN Journal (2019). DOI:https://doi.org/10.2139/ssrn.3446944

[5] Alan Alexander, Megan McGill, Anna Tarasova, Cara Ferreira, and Delphine Zurkiya. 2019. Scanning the Future of Medical Imaging. Journal of the American College of Radiology 16, 4 (April 2019), 501–507. DOI:https://doi.org/10.1016/j.jacr.2018.09.050

[6] Ilkay Altintas, Oscar Barney, and Efrat Jaeger-Frank. 2006. Provenance Collection Support in the Kepler Scientific Workflow System. In Provenance and Annotation of Data (Lecture Notes in Computer Science), Springer, Berlin, Heidelberg, 118–132. DOI:https://doi.org/10.1007/11890850_14

[7] Anand Murali. 2019. How India's data labellers are powering the global AI race. FactorDaily. Retrieved August 24, 2021 from https://archive.factordaily.com/indian-data-labellers-powering-the-global-ai-race/

[8] Marcus A. Badgeley, John R. Zech, Luke Oakden-Rayner, Benjamin S. Glicksberg, Manway Liu, William Gale, Michael V. McConnell, Bethany Percha, Thomas M. Snyder, and Joel T. Dudley. 2019. Deep learning predicts hip fracture using confounding patient and healthcare variables. npj Digit. Med. 2, 1 (December 2019), 31. DOI:https://doi.org/10.1038/s41746-019-0105-1

[9] Khalid Belhajjame, Reza B'Far, James Cheney, Sam Coppens, Stephen Cresswell, Yolanda Gil, Paul Groth, Graham Klyne, Timothy Lebo, Jim McCusker, Simon Miles, James Myers, Satya Sahoo, and Curt Tilmes. 2013. PROV-DM: The PROV Data Model. (2013).

[10] Tellen Bennett, Seth Russell, James King, Lisa Schilling, Chan Voong, Nancy Rogers, Bonnie Adrian, Nicholas Bruce, and Debashis Ghosh. 2019. Accuracy of the Epic Sepsis Prediction Model in a Regional Health System. arXiv:1902.07276 [cs, stat] (February 2019). Retrieved September 5, 2021 from http://arxiv.org/abs/1902.07276

[11] Francine Berman, Rob Rutenbar, Brent Hailpern, Henrik Christensen, Susan Davidson, Deborah Estrin, Michael Franklin, Margaret Martonosi, Padma Raghavan, Victoria Stodden, and Alexander S. Szalay. 2018. Realizing the potential of data science. Commun. ACM 61, 4 (March 2018), 67–72. DOI:https://doi.org/10.1145/3188721

[12] Donald J. Berndt, James A. McCart, Dezon K. Finch, and Stephen L. Luther. 2015. A Case Study of Data Quality in Text Mining Clinical Progress Notes. ACM Trans. Manage. Inf. Syst. 6, 1 (April 2015), 1–21. DOI:https://doi.org/10.1145/2669368

[13] Peter Buneman and Susan B Davidson. Data provenance – the foundation of data quality. 8.

[14] Peter Buneman, Sanjeev Khanna, and Tan Wang-Chiew. 2001. Why and Where: A Characterization of Data Provenance. In Database Theory — ICDT 2001, Jan Van den Bussche and Victor Vianu (eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 316–330. DOI:https://doi.org/10.1007/3-540-44503-X_20

[15] Cansu Canca. 2020. Operationalizing AI ethics principles. Commun. ACM 63, 12 (November 2020), 18–21. DOI:https://doi.org/10.1145/3430368

[16] Center for Devices and Radiological Health. 2020. Classify Your Medical Device. FDA. Retrieved September 9, 2021 from https://www.fda.gov/medical-devices/overview-device-regulation/classify-your-medical-device

[17] James Cheney, Laura Chiticariu, and Wang-Chiew Tan. 2007. Provenance in Databases: Why, How, and Where. FNT in Databases 1, 4 (2007), 379–474. DOI:https://doi.org/10.1561/1900000006

[18] Coalition for Critical Technology. 2020. Abolish the #TechToPrisonPipeline. Medium. Retrieved August 24, 2021 from https://medium.com/@CoalitionForCriticalTechnology/abolish-the-techtoprisonpipeline-9b5b14366b16

[19] Enrico Coiera. 2019. The Last Mile: Where Artificial Intelligence Meets Reality. J Med Internet Res 21, 11 (November 2019), e16323. DOI:https://doi.org/10.2196/16323

[20] Alexander D'Amour, Katherine Heller, Dan Moldovan, Ben Adlam, Babak Alipanahi, Alex Beutel, Christina Chen, Jonathan Deaton, Jacob Eisenstein, Matthew D. Hoffman, Farhad Hormozdiari, Neil Houlsby, Shaobo Hou, Ghassen Jerfel, Alan Karthikesalingam, Mario Lucic, Yian Ma, Cory McLean, Diana Mincu, Akinori Mitani, Andrea Montanari, Zachary Nado, Vivek Natarajan, Christopher Nielson, Thomas F. Osborne, Rajiv Raman, Kim Ramasamy, Rory Sayres, Jessica Schrouff, Martin Seneviratne, Shannon Sequeira, Harini Suresh, Victor Veitch, Max Vladymyrov, Xuezhi Wang, Kellie Webster, Steve Yadlowsky, Taedong Yun, Xiaohua Zhai, and D. Sculley. 2020. Underspecification Presents Challenges for Credibility in Modern Machine Learning. arXiv:2011.03395 [cs, stat] (November 2020). Retrieved August 24, 2021 from http://arxiv.org/abs/2011.03395

[21] Virginia Dignum. 2017. Responsible Artificial Intelligence: Designing AI for Human Values. 1 (2017), 9.

[22] Mengnan Du, Ninghao Liu, and Xia Hu. 2019. Techniques for interpretable machine learning. Commun. ACM 63, 1 (December 2019), 68–77. DOI:https://doi.org/10.1145/3359786

[23] Chris Edwards. 2020. Leveraging unlabeled data. Communications of the ACM 63, 6 (2020), 13–14. DOI:https://doi.org/10.1145/3392496

[24] Aaron M. Ellison, Emery R. Boose, Barbara S. Lerner, Elizabeth Fong, and Margo Seltzer. 2020. The End-to-End Provenance Project. Patterns 1, 2 (May 2020), 100016. DOI:https://doi.org/10.1016/j.patter.2020.100016

[25] Jessica Fjeld, Nele Achten, Hannah Hilligoss, Adam Nagy, and Madhulika Srikumar. 2020. Principled Artificial Intelligence: Mapping Consensus in Ethical and Rights-Based Approaches to Principles for AI. SSRN Journal (2020). DOI:https://doi.org/10.2139/ssrn.3518482

[26] Maayan Frid-Adar, Eyal Klang, Michal Amitai, Jacob Goldberger, and Hayit Greenspan. 2018. Synthetic data augmentation using GAN for improved liver lesion classification. In 2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018), IEEE, Washington, DC, 289–293. DOI:https://doi.org/10.1109/ISBI.2018.8363576

[27] Sorelle A. Friedler and Christo Wilson. 2018. Conference on Fairness, Accountability and Transparency. In Proceedings of Machine Learning Research, PMLR, 1–2. Retrieved August 24, 2021 from https://proceedings.mlr.press/v81/friedler18a.html

[28] Runshan Fu, Yan Huang, and Param Vir Singh. 2020. Artificial Intelligence and Algorithmic Bias: Source, Detection, Mitigation, and Implications. Tutorials in Operations Research (November 2020), 39–63. DOI:https://doi.org/10.1287/educ.2020.0215

[29] Yan Gao and Yan Cui. 2020. Deep transfer learning for reducing health care disparities arising from biomedical data inequality. Nat Commun 11, 1 (December 2020), 5131. DOI:https://doi.org/10.1038/s41467-020-18918-3

[30] Salvador García, Julián Luengo, and Francisco Herrera. 2016. Tutorial on practical tips of the most influential data preprocessing algorithms in data mining. Knowledge-Based Systems 98, (April 2016), 1–29. DOI:https://doi.org/10.1016/j.knosys.2015.12.006

[31] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. 2020. Datasheets for Datasets. arXiv:1803.09010 [cs] (March 2020). Retrieved August 24, 2021 from http://arxiv.org/abs/1803.09010

[32] Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A. Wichmann. 2020. Shortcut learning in deep neural networks. Nat Mach Intell 2, 11 (November 2020), 665–673. DOI:https://doi.org/10.1038/s42256-020-00257-z

[33] Gerard George, Martine R. Haas, and Alex Pentland. 2014. Big Data and Management. Academy of Management Journal (April 2014). DOI:https://doi.org/10.5465/amj.2014.4002

[34] Lise Getoor. 2019. Responsible Data Science. In Proceedings of the 2019 International Conference on Management of Data, ACM, Amsterdam Netherlands, 1–1. DOI:https://doi.org/10.1145/3299869.3314117

[35] Milena A. Gianfrancesco, Suzanne Tamang, Jinoos Yazdany, and Gabriela Schmajuk. 2018. Potential Biases in Machine Learning Algorithms Using Electronic Health Record Data. JAMA Intern Med 178, 11 (November 2018), 1544. DOI:https://doi.org/10.1001/jamainternmed.2018.3763

[36] Justin Scott Giboney, Susan A. Brown, Paul Benjamin Lowry, and Jay F. Nunamaker. 2015. User acceptance of knowledge-based system recommendations: Explanations, arguments, and fit. Decision Support Systems 72, (April 2015), 1–10. DOI:https://doi.org/10.1016/j.dss.2015.02.005

[37] Paul E. Green, Abba M. Krieger, and Yoram (Jerry) Wind. 2001. Thirty Years of Conjoint Analysis: Reflections and Prospects. Interfaces 31, 3 (2001), S56–S73.

[38] Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi. 2018. A Survey of Methods for Explaining Black Box Models. ACM Comput. Surv. 51, 5 (August 2018). DOI:https://doi.org/10.1145/3236009

[39] David Gunning, Mark Stefik, Jaesik Choi, Timothy Miller, Simone Stumpf, and Guang-Zhong Yang. 2019. XAI—Explainable artificial intelligence. Sci. Robot. 4, 37 (December 2019), eaay7120. DOI:https://doi.org/10.1126/scirobotics.aay7120

[40] Shivani Gupta and Atul Gupta. 2019. Dealing with Noise Problem in Machine Learning Data-sets: A Systematic Review.

Procedia Computer Science 161, (January 2019), 466–474. DOI:https://doi.org/10.1016/j.procs.2019.11.146

[41] Jianxing He, Sally L. Baxter, Jie Xu, Jiming Xu, Xingtao Zhou, and Kang Zhang. 2019. The practical implementation of artificial intelligence technologies in medicine. Nat Med 25, 1 (January 2019), 30–36. DOI:https://doi.org/10.1038/s41591-018-0307-0

[42] Jun He and William R. King. 2008. The Role of User Participation in Information Systems Development: Implications from a Meta-Analysis. Journal of Management Information Systems 25, 1 (July 2008), 301–331. DOI:https://doi.org/10.2753/MIS0742-1222250111

[43] Andreas Holzinger, Georg Langs, Helmut Denk, Kurt Zatloukal, and Heimo Müller. 2019. Causability and explainability of artificial intelligence in medicine. WIREs Data Mining and Knowledge Discovery 9, 4 (2019), e1312. DOI:https://doi.org/10.1002/widm.1312

[44] Marijn Janssen, Paul Brous, Elsa Estevez, Luis S. Barbosa, and Tomasz Janowski. 2020. Data governance: Organizing data for trustworthy Artificial Intelligence. Government Information Quarterly 37, 3 (July 2020), 101493. DOI:https://doi.org/10.1016/j.giq.2020.101493

[45] Matthias Jarke and Yannis Vassiliou. Data Warehouse Quality: A Review of the DWQ Project. 14.

[46] Jeffrey Dastin. 2018. Amazon scraps secret AI recruiting tool that showed bias against women. Reuters. Retrieved August 25, 2021 from https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G

[47] Jeremy Hsu. 2020. AI Recruiting Tools Aim to Reduce Bias in the Hiring Process. IEEE Spectrum. Retrieved August 25, 2021 from https://spectrum.ieee.org/ai-tools-bias-hiring

[48] Anna Jobin, Marcello Ienca, and Effy Vayena. 2019. Artificial Intelligence: the global landscape of ethics guidelines. Nat Mach Intell 1, 9 (September 2019), 389–399. DOI:https://doi.org/10.1038/s42256-019-0088-2

[49] Deborah G. Johnson. 2015. Technology with No Human Responsibility? J Bus Ethics 127, 4 (April 2015), 707–715. DOI:https://doi.org/10.1007/s10551-014-2180-1

[50] Jonathan Vanian. 2021. This hot startup is now valued at $1 billion for its A.I. skills. Fortune. Retrieved September 9, 2021 from https://fortune.com/2021/08/09/snorkel-ai-funding-data-labeling-startup/

[51] Gerald C. Kane. 2011. A multimethod study of information quality in wiki collaboration. ACM Trans. Manage. Inf. Syst. 2, 1 (March 2011), 1–16. DOI:https://doi.org/10.1145/1929916.1929920

[52] Katherine Anne Long. New Amazon data shows Black, Latino and female employees are underrepresented in best-paid jobs. The Seattle Times. Retrieved August 25, 2021 from https://www.seattletimes.com/business/amazon/new-amazon-data-shows-black-latino-and-female-employees-are-underrepresented-in-best-paid-jobs/

[53] Christopher J. Kelly, Alan Karthikesalingam, Mustafa Suleyman, Greg Corrado, and Dominic King. 2019. Key challenges for delivering clinical impact with artificial intelligence. BMC Med 17, 1 (December 2019), 195. DOI:https://doi.org/10.1186/s12916-019-1426-2

[54] Derek Khanna. 2013. We Need More Women in Tech: The Data Prove It. The Atlantic. Retrieved September 5, 2021 from https://www.theatlantic.com/technology/archive/2013/10/we-need-more-women-in-tech-the-data-prove-it/280964/

[55] Sebastian Kiefer. 2021. CaSE: Explaining Text Classifications by Fusion of Local Surrogate Explanation Models with Contextual and Semantic Knowledge. Information Fusion 77, (2021), 184–195. DOI:https://doi.org/10.1016/j.inffus.2021.07.014

[56] Jon Kleinberg and Sendhil Mullainathan. 2019. Simplicity Creates Inequity: Implications for Fairness, Stereotypes, and Interpretability. arXiv:1809.04578 [cs, stat] (June 2019). Retrieved September 9, 2021 from http://arxiv.org/abs/1809.04578

[57] Robert W. Kolb (Ed.). 2010. Lessons from the financial crisis: causes, consequences, and our economic future. John Wiley & Sons, Hoboken, N.J.

[58] Himabindu Lakkaraju, Ece Kamar, Rich Caruana, and Jure Leskovec. 2019. Faithful and Customizable Explanations of Black Box Models. In Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society, ACM, Honolulu HI USA, 131–138. DOI:https://doi.org/10.1145/3306618.3314229

[59] Alexandra L'Heureux, Katarina Grolinger, Hany F. Elyamany, and Miriam A. M. Capretz. 2017. Machine Learning With Big Data: Challenges and Approaches. IEEE Access 5, (2017), 7776–7797. DOI:https://doi.org/10.1109/ACCESS.2017.2696365

[60] Xueping Liang, Sachin Shetty, Deepak Tosh, Charles Kamhoua, Kevin Kwiat, and Laurent Njilla. 2017. ProvChain: A Blockchain-Based Data Provenance Architecture in Cloud Environment with Enhanced Privacy and Availability. IEEE, Madrid, Spain, 468–477. DOI:https://doi.org/10.1109/CCGRID.2017.8

[61] Christian Lovis. 2019. Unlocking the Power of Artificial Intelligence and Big Data in Medicine. J Med Internet Res 21, 11 (November 2019), e16607. DOI:https://doi.org/10.2196/16607

[62] Thomas Macaulay. 2020. Flawed Algorithm Used to Determine U.K. Welfare Payments Is "Pushing People Into Poverty." (2020). Retrieved from https://thenextweb.com/neural/2020/09/29/flawed-algorithm-used-to-determine-uk-welfare-payments-is-pushing-people-into-poverty/ http://cacm.acm.org/news/247831-flawed-algorithm-used-to-determine-u-k-welfare-payments-is-pushing-people-into-poverty

[63] Andreas Matthias. 2004. The responsibility gap: Ascribing responsibility for the actions of learning automata. Ethics Inf Technol 6, 3 (2004), 175–183. DOI:https://doi.org/10.1007/s10676-004-3422-1

[64] John A. McDermid, Yan Jia, Zoe Porter, and Ibrahim Habli. 2021. Artificial intelligence explainability: the technical and ethical dimensions. Phil. Trans. R. Soc. A. 379, 2207 (October 2021), 20200363. DOI:https://doi.org/10.1098/rsta.2020.0363

[65] Anay Mehrotra and L. Elisa Celis. 2021. Mitigating Bias in Set Selection with Noisy Protected Attributes. In Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT '21), Association for Computing Machinery, New York, NY, USA, 237–248. DOI:https://doi.org/10.1145/3442188.3445887

[66] Xiao-Li Meng. 2018. Statistical paradises and paradoxes in big data (I): Law of large populations, big data paradox, and the 2016 US presidential election. Ann. Appl. Stat. 12, 2 (June 2018). DOI:https://doi.org/10.1214/18-AOAS1161SF

[67] Michael C. Ksiazek. Sepsis Accounts for 1 in 5 Deaths, Leading Cause of Death in Hospitals. The National Law Review. Retrieved September 5, 2021 from https://www.natlawreview.com/article/sepsis-accounts-1-5-deaths-leading-cause-death-hospitals

[68] Kiran-Kumar Muniswamy-Reddy, David Holland, Uri Braun, and Margo Seltzer. 2006. Provenance-Aware Storage Systems.

[69] Shivaramakrishnan Narayan, Martin Gagné, and Reihaneh Safavi-Naini. 2010. Privacy preserving EHR system using attribute-based infrastructure. In Proceedings of the 2010 ACM workshop on Cloud computing security workshop - CCSW '10, ACM Press, Chicago, Illinois, USA, 47. DOI:https://doi.org/10.1145/1866835.1866845

[70] Neha and Payal Pahwa. 2020. Dark Data Analytics Using Blockchain Technology. In Advances in Data Sciences, Security and Applications (Lecture Notes in Electrical Engineering), Springer, Singapore, 467–474. DOI:https://doi.org/10.1007/978-981-15-0372-6_38

[71] B. Nestor, Matthew B. A. McDermott, Geeticka Chauhan, Tristan Naumann, Michael C. Hughes, A. Goldenberg, and M. Ghassemi. 2018. Rethinking clinical prediction: Why machine learning must consider year of care and feature aggregation. ArXiv (2018).

[72] Regina Nuzzo. 2014. Scientific method: Statistical errors. Nature 506, 7487 (February 2014), 150–152. DOI:https://doi.org/10.1038/506150a

[73] Dino Pedreschi, Fosca Giannotti, Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, and Franco Turini. 2019. Meaningful Explanations of Black Box AI Decision Systems. AAAI 33, (July 2019), 9780–9784. DOI:https://doi.org/10.1609/aaai.v33i01.33019780

[74] Yi Qu, Haitao Wu, Ting Liu, and Yue Zhao. 2018. Space Mission Data Provenance Traceability. In 2018 SpaceOps Conference, American Institute of Aeronautics and Astronautics, Marseille, France. DOI:https://doi.org/10.2514/6.2018-2482

[75] B. Ramesh and M. Jarke. 2001. Toward reference models for requirements traceability. IIEEE Trans. Software Eng. 27, 1 (January 2001), 58–93. DOI:https://doi.org/10.1109/32.895989

[76] Alexander J. Ratner, Stephen H. Bach, Henry R. Ehrenberg, and Chris Ré. 2017. Snorkel: Fast Training Set Generation for Information Extraction. In Proceedings of the 2017 ACM International Conference on Management of Data (SIGMOD '17), Association for Computing Machinery, New York, NY, USA, 1683–1686. DOI:https://doi.org/10.1145/3035918.3056442

[77] Russell Jurney. Hand labeling is the past. The future is #NoLabel AI. KDnuggets. Retrieved September 5, 2021 from https://www.kdnuggets.com/hand-labeling-is-the-past-the-future-is-nolabel-ai.html/

[78] Daniel Russo, Paolo Ciancarini, Tommaso Falasconi, and Massimo Tomasi. 2018. A Meta-Model for Information Systems Quality: A Mixed Study of the Financial Sector. ACM Trans. Manage. Inf. Syst. 9, 3 (November 2018), 1–38. DOI:https://doi.org/10.1145/3230713

[79] Thomas L. Saaty. 1988. What is the Analytic Hierarchy Process? In Mathematical Models for Decision Support, Gautam Mitra, Harvey J. Greenberg, Freerk A. Lootsma, Marcel J. Rijkaert and Hans J. Zimmermann (eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 109–121. DOI:https://doi.org/10.1007/978-3-642-83555-1_5

[80] Philipp Schmidt, Felix Biessmann, and Timm Teubner. 2020. Transparency and trust in artificial intelligence systems. Journal of Decision Systems 29, 4 (October 2020), 260–278. DOI:https://doi.org/10.1080/12460125.2020.1819094

[81] James Shaw, Frank Rudzicz, Trevor Jamieson, and Avi Goldfarb. 2019. Artificial Intelligence and the Implementation Challenge. J Med Internet Res 21, 7 (July 2019), e13659. DOI:https://doi.org/10.2196/13659

[82] Donghee Shin. 2021. The effects of explainability and causability on perception, trust, and acceptance: Implications for explainable AI. International Journal of Human-Computer Studies 146, (February 2021), 102551. DOI:https://doi.org/10.1016/j.ijhcs.2020.102551

[83] Donghee Shin and Yong Jin Park. 2019. Role of fairness, accountability, and transparency in algorithmic affordance. Computers in Human Behavior 98, (September 2019), 277–284. DOI:https://doi.org/10.1016/j.chb.2019.04.019

[84] Paul Slovic and Amos Tversky. 1974. Who accepts Savage's axiom? Syst. Res. 19, 6 (November 1974), 368–373. DOI:https://doi.org/10.1002/bs.3830190603

[85] Jiamei Sun, Sebastian Lapuschkin, Wojciech Samek, and Alexander Binder. 2022. Explain and improve: LRP-inference fine-

tuning for image captioning models. Information Fusion 77, (January 2022), 233–246. DOI:https://doi.org/10.1016/j.inffus.2021.07.008

[86] Joel S. Tieder, Matthew Hall, Katherine A. Auger, Paul D. Hain, Karen E. Jerardi, Angela L. Myers, Suraiya S. Rahman, Derek J. Williams, and Samir S. Shah. 2011. Accuracy of Administrative Billing Codes to Detect Urinary Tract Infection Hospitalizations. Pediatrics 128, 2 (August 2011), 323–330. DOI:https://doi.org/10.1542/peds.2010-2064

[87] Erico Tjoa and Cuntai Guan. 2020. A Survey on Explainable Artificial Intelligence (XAI): Toward Medical XAI. IEEE Transactions on Neural Networks and Learning Systems (2020), 1–21. DOI:https://doi.org/10.1109/TNNLS.2020.3027314

[88] Eric J. Topol. 2019. High-performance medicine: the convergence of human and artificial intelligence. Nat Med 25, 1 (January 2019), 44–56. DOI:https://doi.org/10.1038/s41591-018-0300-7

[89] Mark Velasquez and Patrick T. Hester. 2013. An Analysis of Multi-Criteria Decision Making Methods.

[90] Paul Voosen. 2017. How AI detectives are cracking open the black box of deep learning. Science (July 2017). DOI:https://doi.org/10.1126/science.aan7059

[91] Zirui Wang, Zihang Dai, Barnabas Poczos, and Jaime Carbonell. 2019. Characterizing and Avoiding Negative Transfer. In 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, Long Beach, CA, USA, 11285–11294. DOI:https://doi.org/10.1109/CVPR.2019.01155

[92] Maranke Wieringa. 2020. What to account for when accounting for algorithms: a systematic literature review on algorithmic accountability. In Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, ACM, Barcelona Spain, 1–18. DOI:https://doi.org/10.1145/3351095.3372833

[93] Derek J. Williams, Samir S. Shah, Angela Myers, Matthew Hall, Katherine Auger, Mary Ann Queen, Karen E. Jerardi, Lauren McClain, Catherine Wiggleton, and Joel S. Tieder. 2013. Identifying Pediatric Community-Acquired Pneumonia Hospitalizations: Accuracy of Administrative Billing Codes. JAMA Pediatrics 167, 9 (September 2013), 851–858. DOI:https://doi.org/10.1001/jamapediatrics.2013.186

[94] Andrew Wong, Erkin Otles, John P. Donnelly, Andrew Krumm, Jeffrey McCullough, Olivia DeTroyer-Cooley, Justin Pestrue, Marie Phillips, Judy Konye, Carleen Penoza, Muhammad Ghous, and Karandeep Singh. 2021. External Validation of a Widely Implemented Proprietary Sepsis Prediction Model in Hospitalized Patients. JAMA Internal Medicine 181, 8 (August 2021), 1065–1070. DOI:https://doi.org/10.1001/jamainternmed.2021.2626

[95] Kyra Yee, Uthaipon Tantipongpipat, and Shubhanshu Mishra. 2021. Image Cropping on Twitter: Fairness Metrics, their Limitations, and the Importance of Representation, Design, and Agency. arXiv:2105.08667 [cs] (May 2021). Retrieved August 25, 2021 from http://arxiv.org/abs/2105.08667

[96] Zhiqiang Zheng, Balaji Padmanabhan, and Steven O. Kimbrough. 2003. On the Existence and Significance of Data Preprocessing Biases in Web-Usage Mining. INFORMS Journal on Computing 15, 2 (May 2003), 148–170. DOI:https://doi.org/10.1287/ijoc.15.2.148.14449

[97] Liming Zhu, Xiwei Xu, Qinghua Lu, Guido Governatori, and Jon Whittle. 2021. AI and Ethics -- Operationalising Responsible AI. In arXiv:2105.08867 [cs]. Retrieved August 10, 2021 from http://arxiv.org/abs/2105.08867

[98] 2018. What is GDPR, the EU’s new data protection law? GDPR.eu. Retrieved August 25, 2021 from https://gdpr.eu/what-is-gdpr/

[99] Shining a light on dark data. Accenture. Retrieved August 25, 2021 from https://www.accenture.com/us-en/insights/financial-services/technology-advisory-dark-data

[100] Reducing Sepsis Mortality by One-Fifth with Epic. Retrieved September 11, 2021 from https://www.epic.com/epic/post/reducing-sepsis-mortality-epic