Leveraging AI for Data Provenance: Enhancing Tracking and Verification of Data Lineage in FATE Assessment

Swathi Chundru
Team Lead, Motivity Labs Pvt Ltd
Hyderabad, Telangana, India

ABSTRACT

A record of the sources and processing of data, known as data provenance, holds new possibilities in the ever-growing role that artificial intelligence (AI)-based systems play in assisting human decision-making. Fairness, accountability, transparency, and explainability are the four key virtues that responsible AI builds upon to prevent the terrible consequences that might arise from biased AI systems. This work describes current biases and explores potential applications of data provenance to alleviate them, in an effort to spark more research on data provenance that facilitates responsible AI. We start by going over biases resulting from the pre-processing and data origins. Next, we talk about the practice as it is now, the difficulties it faces, and the solutions that have been suggested. In order to create responsible AI-based systems, we give an overview of how our recommendations might help establish data provenance and hence eliminate biases arising from the origins and preprocessing of the data. We wrap up by outlining future study directions in our research agenda.

INTRODUCTION

Users' concerns about the appropriate development and application of data-driven artificial intelligence (AI) algorithms supporting evidence-based decision making are growing as these algorithms are used more often in all economic sectors. Previous investigations have already given an indication of the catastrophic consequences of biased and erroneous AI suggestions. in high-stakes situations, with instances from the legal and medical fields, including mistreated patients, worsened poverty, erroneous arrests, and unfair prison sentences. Due to the increased awareness of the issues brought up by the most recent social justice movements, professional associations [1] and researchers [18,34] have called for the development of strategies that support the establishment of responsible AI.

Quick advancements in data-generating technologies, such social media, mobile devices, and sensors, have made the issues brought on by poor data quality worse and put the creation of responsible AI systems at risk. Data from these technologies is produced in previously unheard-of quantities and varieties. The majority of applications have profited from the rapid expansion of data availability (volume, variety, velocity, veracity, etc.); however, data quality has received little attention, which has a negative impact on the caliber of suggestions made using this data. This study, which is driven by these worries, looks at how data provenance might strengthen data quality and increase the FATE (fairness, accountability, transparency, and explainability) of AI-based systems. We contend that data provenance—a document that details the sources and methods of data processing—can evaluate and enhance the FATE of suggestions made by AI algorithms, hence fostering confidence in them. The ability to describe and track the life of data—that is, its sources, processing, and application—both forward and backward increases trust. Provenance is important, as the food, fashion, and pharmaceutical industries have long acknowledged [14]. It influences customers' choices regarding what to buy and how to utilize a product, as well as helping to identify its provenance.

The concept of responsible AI is fundamentally linked to the larger discussion of AI ethics, which has drawn a lot of interest from researchers lately. Various high-level ethical standards have been identified by scholars to guide the development of AI systems [25, 48, 97]. Fairness, accountability, and transparency have drawn a lot of attention in this research community, despite the lack of a consistent consensus in this regard. Concurrently, there has been a surge in research on explainable AI , with current conversations focusing on how it may reconcile technological and ethical issues . Explainability in AI enables consumers and professionals to explore and comprehend AI's internal workings. enabling them to recognize such prejudices. In order to bridge these two viewpoints, we concentrate on four crucial and connected traits of responsible AI: FATE. While research on other AI-based system features, such privacy and agency, is still ongoing, our attention is on how FATE may assist companies in identifying and reducing the detrimental effects of biases present in their data. We go over how conflicts between the various FATE traits could arise, how organizations might handle them, and what areas still require further study.

The majority of respectable AI researchers and practitioners today have placed a strong emphasis on algorithm quality.

The suggestions or results of an algorithm, however, are also highly dependent on the inputs of representations, structures, and high-quality data. In this investigation, In the creation of responsible AI systems, we place a strong emphasis on data provenance, which is a crucial component of data quality [13]. Data provenance, for instance, can assist in revealing issues with data quality pertaining to labor-intensive data labeling, which is frequently carried out by untrained personnel [7] and is otherwise hidden. This is especially concerning because, as noted previously, the outputs or recommendations of AI algorithms are frequently utilized as inputs for other AI algorithms, which exacerbates the issue. An algorithm may employ, for instance, the classification of a radiological scan as benign or malignant as an input to another algorithm that generates a risk score for patient readmission. When faced with such circumstances, data provenance can assist in determining the reasons behind the AI algorithm's subpar performance, enhance its interpretability, or reveal that its apparent satisfactory performance was attained for erroneous reasons (e.g., the system was learning from the radiologist's circle on the scan rather than the scan's actual data when identifying a malignant tumor). Data provenance can help to alleviate these issues and ease FATE assessments by providing information about the data's origin and processing [14] (see Table 1).

One major issue with AI-based systems that support important choices is the absence of data provenance. Establishing data provenance can benefit businesses in the long run by fostering trust in the established system and its suggestions, even though it may result in higher short-term expenditures. Our research specifically aims to answer the following query: What is the impact of data provenance on the four interconnected aspects of responsible AI, namely transparency, accountability, explainability, and fairness?

The study examines biases associated with data origins and pre-processing, talks about the state of practice today and the difficulties that come with it, and offers solutions.

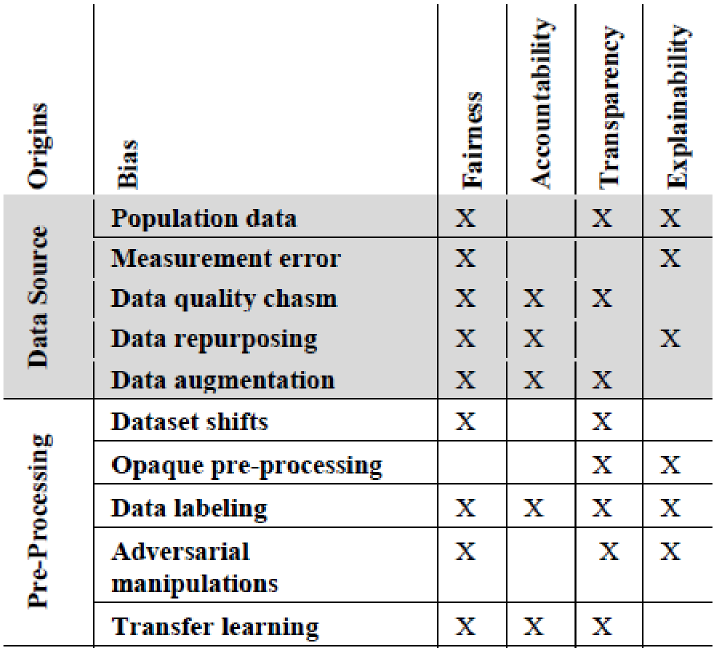
In order to achieve responsible AI-based systems, our guidelines aim to help establish data provenance and eliminate biases arising from the origins and pre-processing of the data.

In the sections that follow, we go over some of the major biases—like systematic distortions [3]—that arise when AI-based systems are developed and deployed without following proper data provenance procedures. In addition, we offer three main suggestions for proving the provenance of data to improve the FATE of AI-based systems. We address model applications for responsible AI and put forth a data provenance system. We outline some avenues for future research before we wrap up.

TWO DATA SOURCES BIASES IN SYSTEMS BASED ON AI

We concentrate on the data's origins and pre-processing rather than the algorithm that uses the data as inputs, in contrast to most previous research, which has concentrated on biases arising from algorithms (e.g., [28,35]). In order to train and construct AI-based systems, original data were frequently gathered from data sources. Biases may also be introduced by data pre-processing, which typically entails data integration, cleaning, normalization, and transformation after data collection. We list five types of potential biases that could come from the sources of the data as well as five categories of biases that could be added during the pre-processing stage of the data. For instance, bias may exist in the data itself due to methods of measurement or sampling. The effects of each bias on the FATE properties of AI-based systems vary.

Table 2: Synopsis of Data Biases' Impact 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	

Biases at the Sources of the Data

The population data, measurement inaccuracy, data quality gap, data repurposing, and data augmentation are the five main ways that biases in the data sources might occur. We outline their ramifications for the FATE attributes (a summary can be found in Table 2).

population information. Sampling the appropriate data to achieve representativeness is crucial in any data science effort. However, developers frequently depend on access to unique data in order to design and execute strong AI-based systems. For example, the medical records of over 11 million patients from eight countries are included in the data offered by initiatives like Big Medilytics. Further data that is representative of the new context is needed for the retraining or recalibration of AI-based systems built with such unique data to different contexts for the same objective.

However, due to the considerable difficulties in gathering the required extra data, AI-based systems are frequently implemented in new contexts without retraining or recalibration. Any variations in the frequency and type of events in these datasets will lead to subpar performance, for instance, when an algorithm trained on data from one population is used to provide predictions on another [19]. Transparency about the provenance of the data is impacted when data collecting systems impose selection bias or overlook mismatches between the training and target populations. Furthermore, the AI system's spurious correlations and shortcut learning—that is, decision rules that perform well based on training data due to spurious phenomena —would produce unfair and erroneous recommendations [20] that will cast doubt on plausible explanations.

mistake in measurement. No matter how well-designed a study or measurement tool is, errors will always occur. Because the outputs of AI applications are always sensitive to probability, Bayesian statistics is widely used in fields like business and medical.

Nevertheless, the uncertainty of the input variables arising from pre-processing or the measurement itself is frequently disregarded in AI systems. Without paying special attention to and exercising caution regarding any inaccuracies, an AI-based system trained with such data may produce a model that performs badly. As a result, when an AI-based system learns to adapt against mistake, its precision may be exaggerated. Problematic results would ensue from the recommendations that would be produced, if not outright inaccurate then at least skewed. A user can see these shortcomings and make the necessary corrections if the system offers matching explanations [19].

A gap in data quality. The absence of sufficient-quality data in environments where the AI system is employed presents another difficulty. Even if the data could appear homogeneous at first glance, a closer inspection may reveal otherwise. For instance, an AI program with access to the most recent computed tomography (CT) images may be able to produce predictions with a higher degree of accuracy. Retraining the AI-based system with CT scans from outdated equipment that produces lower-quality scans is likely to result in erroneous suggestions. Here, the AI-based system performed worse because it was trained using fine-grained data that were later rendered unavailable, as opposed to measurement error, where the system learnt to forecast based on errors. This offers several difficulties along the FATE traits. Inadequate performance may result in suggestions that are not as good as they may be, and depending on the original degree of transparency, issues of accountability between the system provider and developer may come up. Mitigating this issue is made easier by establishing transparency on the origins of the training data and the data utilized to generate the recommendations.

Repurposing data. Data collection procedures bring biases and abuse in addition to sampling-related biases.

AI system development procedures nowadays are very different from those used in the past when it comes to data acquisition. The conventional approach is to gather information with a specific goal in mind. For instance, experimental data will be gathered during a clinical study of a medication used to treat COVID-19 in order to evaluate the drug's negative effects.

Repurposing data, however, is standard practice in AI-based systems. An AI-based system may, for instance, utilize a blood test result that was recorded in a patient's electronic medical record to detect another condition in addition to the original one. This might be a problem that undermines the algorithm's accountability feature. For instance, the quality of data from medical pictures may be enough for a given use, such as the identification of strokes, but it might not be adequate for other uses, such as the discovery of novel disease markers [5]. The act of repurposing data introduces uncertainty regarding the data and its source, hence impeding the ability to definitively identify the individual or organization responsible for any inaccurate suggestions.

data enhancement. Data augmentation is the process of adding artificially generated data or slightly altered copies of the existing dataset to increase its size when the present dataset is insufficient for the desired computations (e.g., translation, rotation, flip, or scale). For example, when training a generative adversarial network (GAN), augmented data are produced by rotating, translating, and scaling a previous dataset on liver lesions. These changes, along with the artificially produced data, have the potential to exacerbate preexisting biases in the dataset and conceal the shortcomings of the information gathered.

Certain AI systems just use synthetic data. AI systems, for instance, have been designed to use solely simulation data to control robot arms and build bridges. In situations where there is limited input and manually labeled data, simulations can generate valuable data that can be used for learning. However, academics have proposed that AI systems perform better without artificial data additions because deep learning may approach issues more naturally by focusing on patterns in the core data.

Thus, additional issues with the fairness and accountability of AI algorithms arise from data augmentation and the usage of simulation data. Data augmentation reduces transparency and makes it more difficult to pinpoint the source of an inaccurate suggestion since it magnifies preexisting biases and produces uncertainty about the data's true representativeness.

Pre-processing Data Biases

Errors that add biases into data processing can occur from transfer learning, adversarial manipulation, dataset changes, opaque pre-processing, data labeling, and data augmentation.

Changes in the dataset. The non-stationary nature of the environment and people, which provide all of the input data for AI-based systems, is an easily overlooked feature. A key predictor of a particular disease at one moment in time, for instance, may become less or more significant at a later time due to advancements in the quality of care that is accessible, as a result of a data shift. For example, changes in hospital operation practices confuse many forecasts made with the Medical Information Mart for Intensive Care dataset. Taking time into account as a crucial variable reveals dataset adjustments brought about by evolving methods, which in turn generate notable modifications in the observed data. The system's performance declines and the algorithm's attributes of fairness, transparency, and explainability are impacted unless this data shift is found and the AI algorithm is retrained or recalibrated.

Inaccurate recommendations brought on by poor performance may have a detrimental effect on users. In the event that the sources of the data and any ensuing environmental modifications remain opaque, the inferred explanations will be, at minimum, warped.

opaque initial processing. Black boxes are a common term used to describe AI-based technologies [2]. Even if some AI-based systems make correct predictions, it's still unclear why they think the way they do.

Deep neural networks are examples of algorithms with intrinsic obscurity, making it challenging to understand the precise patterns being learned . For instance, the scanner model and scans designated as "urgent" confused an algorithm in a study that detected hip fractures [8]. As a result, it can be challenging to evaluate the possible biases brought about by feeding the results of an opaque algorithm into another AI-based system. The transparency and explainability of AI-based system suggestions are restricted by opaque preparation. Uncertainty about the data used to train the system makes it more challenging to recognize and evaluate confusing indications, which prevents users from gaining pertinent insights. On the other hand, professionals can validate the model and its recommendations by developing reasons for the recommendations. An expert can assess, enhance, and adjust the model with the aid of several explanations (such as feature extraction, pre-defined models, and sensitivity).

Data labeling: While the term "data quality chasm" describes data that may seem comparable but actually have distinct qualities, "data labeling" also presents a problem because it is frequently opaque how labels are identified and developed. Supervised learning, like the classification of medical images, is associated with data labeling. supervised algorithms use the outcome labels during the training phase. Although data labeling is becoming more automated (e.g., with loose monitoring) [50,76], labeling is still a labor-intensive activity that is commonly completed by inexperienced or unskilled ghost workers or through crowdsourcing platforms [7].

Due to the inherent bias in the training data, incorrect labels lead to inaccurate or unjust suggestions and explanations produced by AI-based systems. The AI algorithm's fairness, transparency, and explainability are all impacted by this bias. Due to the possibility of errors and the introduction of social biases into the data, untrained or underqualified ghost workers compromise the objectivity of the data. Since these are bad business practices, companies rarely reveal them, which has a detrimental effect on transparency. Although there are biases introduced by these business methods, it is more difficult for the user and the expert to gain from explanations when they are concealed from customers.

Some academics believe that the reliance on labeled data may even be detrimental to the development of effective AI, as the majority of available data are unlabeled and typically need significant financial resources to label. There has been a significant upsurge in the use of AI for data tagging automatically. The concept is straightforward. In the process of developing AI systems, labeling is frequently a bottleneck operation. Therefore, we might utilize machine learning (ML) to extrapolate the labels. A labeling machine learning algorithm can be trained on a small collection of readily available or attainable labels before being applied to a larger dataset.

Although this lessens the effort required for physical labor, it may also exacerbate biases that were already present in the smaller sample, resulting in inaccurate to unfair or incorrect advice and justifications. manipulation by adversaries. Small changes in the data input might occasionally result in major alterations in the output since AI-based systems rely their models on subtle variances in the data . As a result, systems built on AI may be vulnerable to hostile manipulation. For example, additional adversarial noise or seemingly little changes in the data can lead to the incorrect diagnosis of benign moles in photos as cancer.

These alterations might be unintended, like when a user inadvertently flips an image used as an input, or malicious, such when an attacker modifies the algorithm's input to trick it. Lacking adequate transparency in the preprocessing of the data, It is challenging to recognize this possible danger in an otherwise successful approach. These ostensibly little adjustments may have profoundly divergent effects, making it challenging to explain the suggestions and perhaps even erroneous.

Transfer knowledge. When an AI-based system is developed, we might apply the algorithm to other issues of a similar nature. Specifically, the knowledge gleaned from one AI-based system helps a new one. To increase the sample efficiency for a reinforcement learning agent, for instance, radiographic features in images might be encoded using a pre-trained model before final re-training [8].

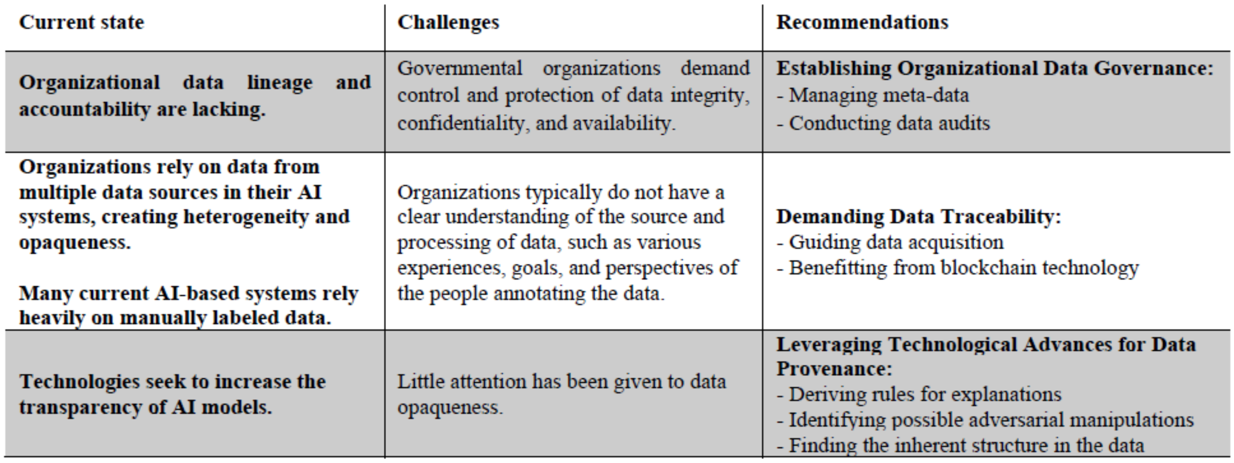
Additionally, transfer learning can enhance the accuracy of AI systems in cancer prediction for ethnic groups for which there is a dearth of available data. However, only when the original task and the new task are closely related can transfer learning take place. If not, transfer learning results in bias introduction and a decline in performance. Transfer learning prevents clear responsibility because it also makes the recommendations made by the AI-based system more ambiguous. Transfer learning should therefore be made transparent to the user, as it would otherwise increase the opaqueness of the system.

THREE SUGGESTIONS FOR USING DATA PROVENANCE

When building responsible AI-based systems that address the FATE features, companies must establish data provenance in light of the significance of minimizing data-induced biases emanating from data sources and data pre-processing. To improve responsible AI's FATE attributes, we provide a data provenance system (Figure 1).

Establishing organizational data governance, requiring data traceability, and utilizing technology advancements like explainable AI are the three main areas on which organizations can concentrate. The issues of the present and the future are outlined below, along with concrete suggestions and an explanation of how they will improve the particular attributes of responsible AI (refer to Table 3).

Table 3: Summary of the situation as of right now, difficulties, and suggestions



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

Putting in Place Organizational Data Management

To ensure control and protection of data integrity, confidentiality, and availability, a number of governmental entities have introduced directives, rules, and regulations. The General Data Protection Regulation (GDPR) of the EU and the Health Insurance Portability and Accountability Act (HIPAA) of the US are two examples. Unfortunately, master data management—a collection of procedures pertaining to the who, what, and where of communications, events, and business transactions—is frequently the exclusive focus of existing data governance methods.

Organizations seem to follow their rivals' lead far too frequently instead of taking the initiative and directing the situation. As an illustration, A lot of businesses are still aiming to be data-driven. However, once they do, they discover that data governance is not adequately considered while AI systems are being developed, which leads to more difficulties.

Establishing organizational data governance procedures that ensure data accountability and lineage is necessary for organizations. In addition to helping businesses comply with ever-tougher regulations, this would give them a comprehensive view of their data assets. Organizations must manage their meta-data in particular and carry out data audits to address the organizational issues brought on by insufficient data governance. Certain organizations may find these objectives to be at odds.

For instance, data privacy aims to prevent people from being identified or connected to such information—often through personally identifying information. On the other side, data lineage describes how the data's sources and subsequent processing may be seen. Both ideas are at odds if people are the ones who collect the data and process it further. In order to resolve this issue, a firm must improve responsible AI while adhering to privacy regulations like the GDPR. For instance, an organization might permit the tracing of personal data only in accordance with certain legal requirements. It is imperative for organizations to utilize privacy-preserving strategies, like federated learning, in order to facilitate the secure exchange of identifiable information or models among various entities.

Taking care of metadata. Meta-data are extensive descriptions of data found in a data source that include information about the data itself. An organization's data can be maintained with the use of metadata, which guarantees the prompt, accurate, and efficient retrieval of the necessary information. It also aids in making sure that procedures and actions are transparently and independently verified in their documentation. Organizations often handle meta-data using two methods: data curation and data cataloging. Information about the data, including the reasons behind selecting a particular data source, the parties involved, and the material included therein, is kept in a data catalog. A datasheet may also contain documentation of this kind.

To further these efforts, businesses must to set up distinct procedures and roles for data curation. Data curation locates and makes use of the organization's data while assisting in evaluating the FATE of system recommendations. Organizations, for instance, can see and cluster data annotations to determine representation and related limits. The detection of discriminatory associations between features, labels, and groups is made easier by these annotations.

In general, controlling meta-data through data curation and catalogs increases the value of already-existing data by promoting transparency and lowers expenses by preventing pointless data collection. It's also necessary to have distinct accountability for the various data sources while managing meta-data. By reducing the degree to which data diverge from the goals of responsible AI, meta-data assist companies in reaping the benefits of transformation, weighting, and sampling approaches [4], so assisting in ensuring the suggestions' fairness.

carrying out audits of data. Another way to achieve data provenance through data governance is to improve an organization's data auditing capability. The practice of determining if data are appropriate for a certain use is known as data auditing. Organizations should evaluate the data utilized in their systems through data audits, just as they evaluate and audit other areas of their company operations, in light of the recent rise in regulatory obligations. Data audits assist in identifying possible biases in data processing and the effects they may have. Data audits contribute to the improvement of AI-based systems' accountability and fairness by providing a reasonable and appropriate guarantee of authenticity and trustworthiness. This benefits other firms that aim to behave responsibly in addition to high-reliability organizations that must make important judgments. Data audits include data profiling, which evaluates the risks involved in data integration as well as the quality and availability of data, and impact analysis, which evaluates the effects of subpar data on profitability and performance.

As separate services are combined into bigger systems, data audits become more crucial. By guaranteeing a good fit between the data and their use, data audits improve the fairness of AI systems. It is also necessary to establish explicit accountability for the proper handling of data while conducting data audits. Data audits not only verify the correctness of the data but also identify data silos and places where additional depth and/or breadth of data are required in order for the AI-based system to generate meaningful recommendations. Simplifying the audit process, a data provenance record might list the entities responsible for collecting and processing the data for the relevant dataset. Data provenance records also aid in deciphering the pre-processing and origins of the data, increasing transparency.

Requiring Information Traceability

Using various data sources and processing techniques has consequences that managers should be aware of, particularly when trying to create systems that are equitable and transparent. Managers are becoming more conscious of the significance of data traceability. For instance, it often takes Walmart six days and fourteen hours to determine where a farm product comes from. However, comprehensive data traceability may be established in just 2.2 seconds when supply chain data are kept on a blockchain. Thus, platform To increase business and decision-making efficiency, suppliers must strengthen the traceability feature of data provenance.

Improved traceability raises general transparency and offers additional details about the historical background of data. An intermediate representation of the original data that encodes the responsible AI goals, such fairness, can be created thanks to transparency [4]. As a result, organizations enhance the fairness of their systems and reduce biases arising from data sources. Enforcing data traceability may involve using blockchain technology and directing data collection.

directing the collection of data. A lot of AI-based technologies in use today require manually labeled data. Even though automated labeling techniques have been used more frequently recently, manual labeling is still necessary. Either all of the dataset or just a portion of the datapoints for future extrapolation are subject to manual labeling. In any case, companies cannot take into account the major influence on data quality if they do not have a comprehensive awareness of the diverse experiences, objectives, and viewpoints of the individuals annotating the data. Data traceability should be taken into account while developing procurement guidelines by organizations. Managers must, for instance, insist on openness about the source and caliber of data when obtaining external training datasets. A data provenance record reveals the often-hidden history of the data by identifying the actual source and any later processing.

Through improved data traceability, a collection of tools, including R packages, have been developed by recent end-to-end provenance efforts to help businesses achieve data provenance.

Additionally, some of the data that the system was trained on might not have had professional labeling, while other data might have come from data brokers—companies that gather data with the intention of reselling it. Knowing where the data came from and how it was obtained is essential to making sure it was done legally and ethically (e.g., with informed consent). Enforcing traceability (e.g., via a data provenance record) promotes openness and assists businesses in identifying the responsible parties for reducing risks associated with the use of suggestions from AI-based systems.

For example, an organization should include the dataset's descriptive statistics in its data provenance records so that consumers can determine whether there is a chance of prejudice. Users can assess the AI-based system's suggestions to rectify, lessen, and prevent prejudice in the future by adjusting the algorithm, the input data, or the prediction process based on these statistics [4].Because of this, the user is more likely to think that the AI-based system's recommendations are reasonable.

Data provenance also improves transparency because it pertains to a record of the data's beginnings and subsequent processing [9]. To create a data information sheet that offers information on the key factors influencing the recommendations of an AI-based system, for instance, data provenance is required. Because of this, data provenance gives consumers a foundational understanding of the data and how they were processed [17] before the AI black box uses them. The appropriateness and relevance of the data used to train the system can be verified by the user .

gaining advantages from blockchain technology. Data provenance based on blockchain technology is a promising way to improve data traceability in responsible artificial intelligence. Blockchains have the ability to store data objects' histories and meta-data. Blockchains' key features—transparency and auditability, for example—allow meta-data to be traceable and secure, which is essential for data accountability. Immutability of data in a blockchain also improves the recommendations' perceived fairness. Numerous blockchain-based data provenance designs have been proposed, including ProvChain and Lineage Chain.

Blockchain technology has also been applied to dark data handling. Dark data is information that companies gather but do not use to its full potential. Blockchain, a distributed ledger that is secure, has the potential to increase the data's value and provide more effective and transparent outcomes.

Organizations are increasingly adopting a consumer-centric strategy, which is supported by more openness. Patient-centered care, for instance, is defined in the healthcare industry as respecting and accommodating each patient's unique requirements, beliefs, and preferences. To do this, health IT systems must place a high priority on data provenance and patient privacy. Patients who have greater access to information are better able to ask questions about diagnosis and recommendations because they are more informed . The standard of medical care is raised by this exchange. Additionally, it increases patients' trust in the quality of the care they receive. Decisions and suggestions made by healthcare organizations are more transparent when they take data provenance in electronic health records into consideration.

Making the Most of Technology Developments for Data Provenance

Data provenance is crucial for comprehending the suggestions made by AI-based systems, as many of these systems are opaque. Explainable artificial intelligence (XAI) methodologies, GANs, and deep learning with improvements in small data techniques are examples of recent technological advancements.

obtaining justifications. The conventional bounds imposed by trade-offs between the accuracy and interpretability of AI systems' suggestions are pushed by XAI techniques as LIME, LORE, and Anchor . More recently, XAI systems have made it possible for users to alter model features and personalize the model explanation, as well as comprehend the key elements that influence the results .

While explainable AI techniques aim to make AI models more transparent, data opaqueness has received less attention. Data provenance offers an additional viewpoint on transparency for the user [6] by outlining the origin and additional processing of the data that powers an AI system. Data provenance contributes more information to the XAI's explanations. systems. For instance, applying the idea of data provenance to AI algorithms makes it easier to record, using both local and global explanations, the data processing that an AI system performs. While a local explanation offers transparency for a specific recommendation (e.g., addressing the question of why the AI makes a specific recommendation for a given patient), 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). Explainable AI, for instance, can help patients and healthcare professionals better comprehend the critical elements that influence an algorithm's recommendations on a certain diagnosis or course of treatment, which will increase accountability among those providing and receiving care. As a result, we recommend that businesses work to fully utilize the most recent XAI-related technological advancements.

To provide easily understood explanations of AI-based recommendations, we specifically recommend that organizations make use of already-existing XAI methods and techniques, like gradient-based explanations and layer-wise relevance propagation, along with supporting architectural frameworks like CaSE . For instance, XAI techniques identify the aspects that have the greatest influence on a suggestion or present cut-off values that result in the desired conclusion in order to develop rules that explain how a recommendation was made. Users can find new patterns in the data and gain a better understanding of the AI system's behavior with the aid of such explanations.

Nevertheless, earlier research also raises the possibility of a contradiction between explainability and other FATE characteristics. For instance, explainability and fairness have trade-offs . Explainability aims to make AI-based systems less complicated so that people can understand them, but this simplification comes with a cost that could introduce new biases. Organizations can handle these conflicts by directing and prioritizing various qualities through the use of multi-criteria decision-making techniques (see for an overview). In a certain situation, one trait may be more significant than another. For instance, explainability may be one technique to improve a system's transparency and foster greater trust if adoption and usage of the system are issues . Users are less likely to reject the system in companies that allow users to contribute to the system's evolution in order to solve potential fairness problems.

The inexplicability of AI prediction results may stem from biases present in the data as well as the opaque nature of algorithms. We recommend focusing more on how data provenance can improve the explainability of results, even though the majority of research focuses on algorithm explainability. Organizations support autonomous decision making, mistake detection, bias minimization, and justice preservation by enabling meaningful human interaction with the system and improving the explainability of AI-based systems [15].

Handling noisy data. "Noise within the data" refers to the existence of irrelevant and nonsensical data. Researchers have come a long way in handling noisy data, which is useful for businesses. There is a differentiation established about the noise's relationship to either target attributes (also known as class noise) or predictive attributes (also known as attributed noise). A variety of methods exist for locating and managing noise in the data. A new systematic review offers a comprehensive summary of the state of the art concerning issues brought on by noisy data in AI-based systems.

In order to arrive at fair suggestions, it is imperative to effectively handle noisy data. In actuality, attempting to attain fairness without taking into account the noise present in a particular dataset may backfire. For instance, a previous study looked into the process of denoising data during subset selection using noise models. To choose a subset of data from an already-existing, bigger data set, researchers used noise models.

The objective was to produce a fair dataset using noisy race data such that the sub-dataset accounts for race. The research highlights the inadvertent consequences of neglecting noise, as it diminishes the impartiality of the ensuing subset selection process.

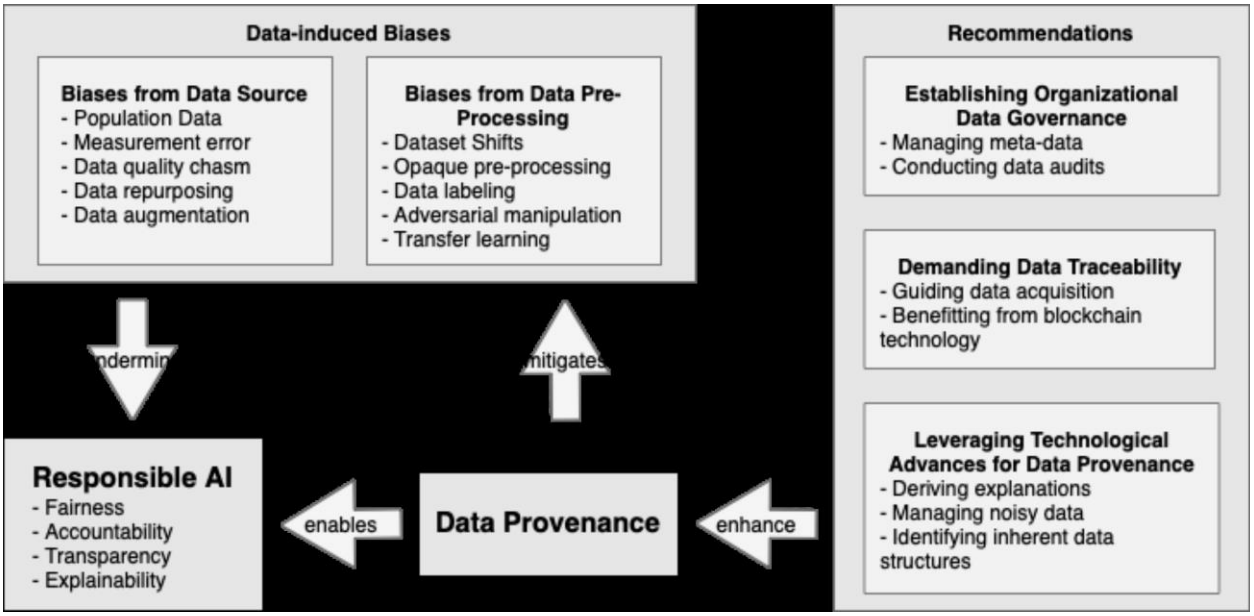
There are various methods for dealing with noise in data. Organizations can, for instance, modify the data—a process known as data polishing—or utilize filtering algorithms to find and eliminate noise. Responding to class noise differs significantly from responding to attribute noise in that organizations must take into account relabeling for class noise while employing data imputation for attribute noise.

Using GANs, which are collections of neural networks that aim to produce new data with properties similar to the training data, is a related method. GANs should be used by organizations to spot potential adversarial manipulations and lessen their detrimental effects. GANs are employed, for instance, in image-to-image translations, such as converting noisy low-dose CT scans into regular-dose CT scans. In this instance, a discriminator attempts to discern between fake and actual regular-dose scans, while a generator network converts the low-dose scan into a regular-dose scan. Consequently, there is less noise in the translation of one image to another.

Recognizing innate data structures. In order to identify an innate structure in the data of their AI systems, deep learning for text, audio, and video recognition frequently entails completing a pre-text task. Self-supervised learning is the pre-text task, and its goal is to produce a usable feature representation for the downstream work [12]. Pre-text challenges could compel machine learning models to break down data in order to improve explainability. For instance, the Facebook AI Research team increases the amount of unlabeled data used in its picture classifier by combining clustering and training using rotated photos. Following this pre-text task processing, typical labeled data is used in the second training stage to provide results that are comprehensible.

Additionally, enterprises can enhance the performance of AI-based systems with the help of advancements in small data techniques. Large data sets are necessary for many AI-based systems, yet some of the most valuable datasets are scarce and only available in limited numbers . For instance, medical personnel like radiologists and doctors are frequently needed to label data in order to apply AI in the field of medicine. To accurately diagnose the presence or absence of lung cancer in an imaging scan, a radiologist's assessment is required.

Large dataset development is difficult since medical experts have limited and expensive time, and data labeling is a recurring effort. But high-quality human input is what makes AI-based systems able to produce high-quality recommendations.



The framework for data provenance in responsible artificial intelligence is shown in Figure 1.

MODEL UTILIZATION OF THE DATA PROVENANCE STRUCTURE

We examine how to apply our approach using two current cases that illustrate the issues raised by the absence of responsible AI. The use of AI suggestions in healthcare is one recent instance of data provenance issues. The ability of the AI-based system integrated into EPIC , a significant electronic healthcare records system, to predict sepsis—a potentially fatal illness in which the body's reaction to an infection damages its own tissues—was assessed in a recent study. Sepsis is the leading cause of death in US hospitals, thus diagnosing and treating diseases that increase the risk of sepsis is very important. Models for predicting sepsis, like the one offered by EPIC, are widely used.

Nevertheless, the analysis indicates that i) the AI-based system does not operate as promised, ii) significant presumptions that In general, a better comprehension of the data and the behavior of the system aids in assessing how fair the recommendations are. This is significant because, for instance, evidence-based medicine is predicated on strict explainability requirements since sound knowledge of underlying illness mechanisms and appropriate treatments for specific circumstances is necessary for medical decision-making. The application of AI in healthcare is hampered by this lack of knowledge. The potential advantages of AI in healthcare make this a critical problem. underlying the AI system need to be carefully examined, and iii) the high frequency of false positives in the system adds to medical staff members' alert fatigue.

Four significant biases are shown by this case study: data shifts, population bias, transfer learning, and data repurposing. A significant finding of the assessment was the possibility that the data used to create the model had been reused. Instead of using the clinical definition of sepsis, EPIC measured positive sepsis instances based on billing data in order to generate the forecasts. Utilizing billing codes also leads to population bias because sepsis is only present when it is recognized by medical personnel. However, the medical team employed the technology in the hope that it would aid in the early detection of sepsis before medical workers could recognize it. EPIC has responded to the findings by stating that subpar performance might be explained by transfer learning. That is to say, the sepsis prediction model created from data from one environment could not perform well since transfer learning only functions when the source task is strongly related to the new task. in different settings. In contrast to data from the University of

Colorado Hospital [10], the sepsis prediction model employing data from the University of Michigan Hospital may have performed poorly due to biases introduced by transfer learning. Finally, the researchers highlight the necessity of completely retiring outdated models and discuss the possibility of a shift in the dataset as a result of altered sepsis treatment procedures. We recommend that companies utilizing these kinds of prediction models set up organizational governance, perform data audits, and take advantage of XAI-related technology advancements to get the reasons behind the models.

While data auditing is the process of determining whether or not the data is appropriate for a given purpose, organizational data auditing capacity ensures data provenance through data governance. Healthcare companies could assess the data that was used to train the AI system and find potential issues with it by conducting a data audit. In our case, a data audit would enable a medical professional to spot possible mistakes arising from the application of billing codes as a stand-in for the existence of an illness. Nevertheless, billing codes can differ from the medical diagnosis and are utilized in the administrative process (e.g.,).

Billing codes are frequently used in research to identify patients for a subsequent study, thereby focusing on those who are most likely to have a particular sickness or condition (e.g.,).

The need for an institution to be able to audit AI systems has grown as a result of a recent study that revealed a serious lack of openness on the part of AI system providers and a lack of FDA monitoring. Experts in medicine questioned EPIC's lack of clarity and transparency. The developer has revealed relatively little about the creation of the prediction model because the AI system is shielded by intellectual property rights. Although the FDA's control was implicitly trusted by medical experts, a recent study highlights its limitations. The FDA rates medical equipment into three classifications [16], with life support systems being assigned the highest class. Systems that are capable of making decisions on their own, such automated insulin pumps or pacemakers, must adhere to the strictest regulations established by the FDA. Artificial intelligence (AI)-based systems that advise medical professionals (such as a sepsis prediction model) are frequently categorized as class II systems, which are subject to far less FDA regulation. The analysis indicates that, in the EPIC example, not even the decreased oversight was used since, although the system may have been examined at the time of its introduction to the market, subsequent modifications are not subject to additional FDA approval.

Organizations can discover necessary adjustments with the use of recent technological advancements. For instance, explainable AI assists in giving AI developers information and feedback so they can modify the network design or retrain the model to further improve the AI system. Scholars have supported the notion known as "human-in-the-loop" as a means of debiasing AI systems. Here, human-in-the-loop feedback is supported by technological advancements in XAI, which can improve data provenance and the prediction model's transparency. For instance, a physician could doubt the model's ability to accurately forecast sepsis early on, but the majority of the concerns described are not exclusive to the healthcare industry. Another example is the Amazon AI recruitment tool, which has drawn criticism for not adhering to the principles of responsible AI (e.g.,). Amazon created an experimental hiring system that was intended to automatically screen job applicants' resumes and identify the best candidates. Later, Amazon discovered that the AI system was biased negatively toward female candidates and did not select candidates for technical jobs in a gender-neutral manner.

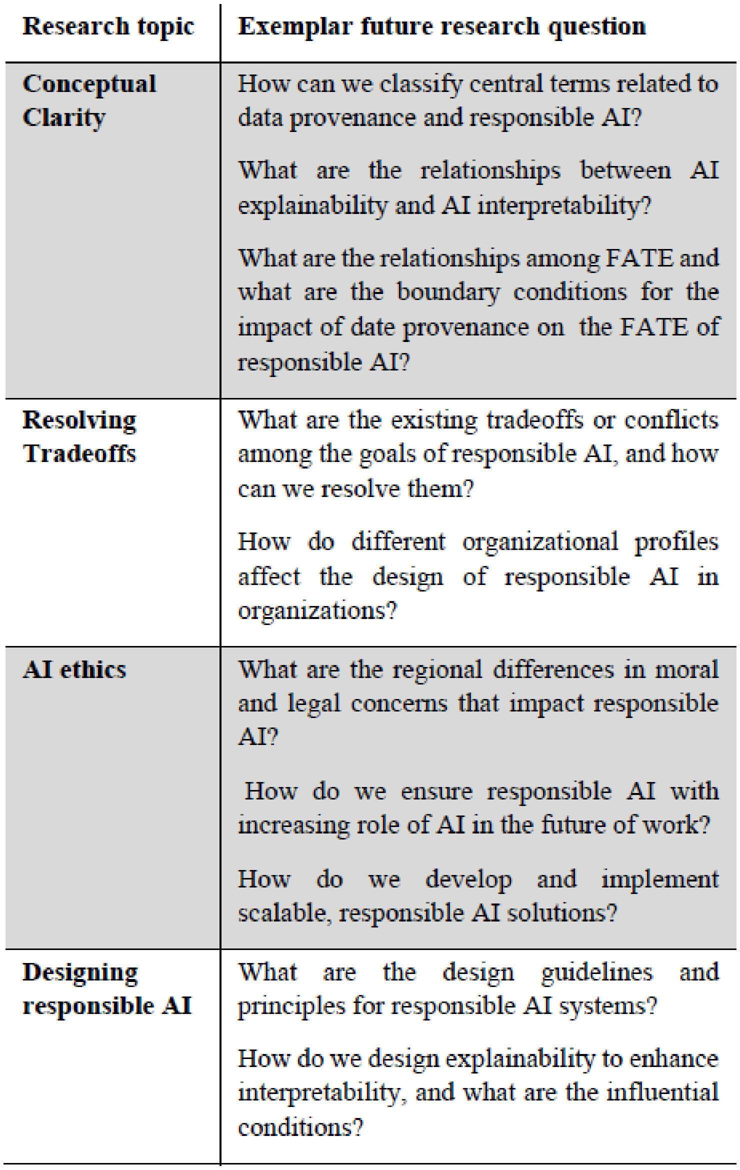
The system was trained using recruitment data that Amazon had collected over the last ten years. It would have been possible to detect the existence of a dataset shift and demographic bias by improving data provenance through an auditing process. It would have specifically brought attention to the fact that female candidates have been treated unfairly in recruiting processes over the last ten years. Additional modifications are required to guarantee responsible AI suggestions. Data auditing, thus, can contribute to a system's increased fairness by proving the provenance of data.

Similarly, firms can assess the AI system with the use of the human-in-the-loop, which has been recommended as a means of debiasing HR recruiting systems. By providing feedback through human-in-the-loop, technological advancements in XAI improve data provenance and lessen the adverse effects of dataset shifts. By using data provenance, XAI improves responsible AI's explainability.

RESEARCH PRIORITIES

AI-based technologies are still being adopted and used by organizations to assist in making evidence-based decisions. Improving the FATE of the AI-based systems that have been put into place is a special emphasis. Three key recommendations for companies are the result of our analysis of data-induced biases and our discussion of how organizations might reduce them by establishing data provenance within their own organizations. However, additional study is required to enhance data provenance techniques, resources, and ethical AI policies. As a result, we formulate suggestions for more study, highlighting four key areas (see Table 4).

Table 4: Model research questions for ethical artificial intelligence



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?

The creation of a distinct nomological network is essential for comprehending the differences between concepts and their relationships, as well as for developing data provenance for responsible AI. Further investigation is required to ascertain the distinctive characteristics of various conceptions and maybe the interchangeability of some concepts. Researchers can find classes with mutually exclusive and collectively exhaustive dimensions by using taxonomy creation techniques. For instance, although they are frequently used interchangeably, explainability and interpretability are really two related but distinct notions; similarly, although they are closely related to data provenance, terminology like data lineage and data pedigree are not the same. More research can be done to comprehend the connections between ontologically disparate notions with improved conceptual clarity.

It's also critical to comprehend the settings in which these partnerships arise. For instance, good recommendations or a fair dataset do not always equate to high transparency. This may aid in the explanation of contradictory findings in the literature. For instance, when it comes to the connection between explainability and transparency, some academics contend that explainability improves transparency, while others contend that explainability is a quality that exists independently of openness. To create a nomological network around data provenance for responsible AI, more investigation is required.

Striking a trade-off. Responsible AI implementation of data provenance may result in tradeoffs or conflicts. Regulations like the GDPR, for instance, mandate that the system protect user privacy, but other requirements—including auditing requirements—call for more traceability. The Twitter cropping algorithm case study illustrates a trade-off between accuracy and speed and prediction error risk. Furthermore, a debate about responsible AI frequently centers on the trade-off between interpretability and accuracy. To recognize these tensions and create appropriate solutions, more investigation is required. For example, certain research techniques like conjoint analysis and the analytic hierarchy process (AHP) methodology can help researchers prioritize various traits or pinpoint crucial combinations of traits in certain contexts.

We propose two key directions to address these problems. First, studies on multi-criteria decision making could be helpful to academics. When making decisions, managers might consider multiple goals that may conflict with one another by consulting prior studies. Before normative advice can be derived from them, they must be extended and evaluated for responsible AI.

Secondly, to offer models for creating responsible AI projects, organizational or AI project profiles could be made. Prioritization can be the consequence of internal organizational culture and values as well as external factors like laws and regulations. For instance, an open and progressive organization might put fairness and openness ahead of worries about accountability. On the other hand, a risk-averse company can prioritize performance and accountability over openness. Similar to this, several initiatives inside a company could require focusing on certain FATE components. Future studies might examine how project-specific and organizational profiles relate to the creation and application of ethical AI systems.

AI morality. Research on innovative technology and its ethical behavior frequently intersects with questions about the fairness of responsible AI . Questions of morality and law are intimately linked to research in ethics. In response to the regional requirements of the judicial system, legal research is frequently carried out at the federal level. In contrast, independent of local needs, new technical obstacles arise throughout the development and deployment of responsible AI-based systems. When it comes to local regulatory regulations, like the GDPR, responsible AI, for instance, raises issues but has the ability to provide easily scalable technology solutions.

The phrase "responsibility gap," which was first used in earlier study , describes a scenario in which artificial entities are utilized to choose a course of action or act independently without human interaction. Since the guidelines they follow are not predetermined at the time of development, no one takes accountability for the activities of the machine. There is a responsibility gap since the ethical and legal structures in place today were not intended for these circumstances . Organizations must frequently pursue several objectives, such as accountability and transparency , in addition to reducing or eliminating the responsibility gap when developing responsible AI systems.

It's unclear how government rules that businesses must abide by relate to various responsible AI objectives, though. Future studies could, for instance, look into whether and how laws like the GDPR in the EU and HIPAA in the US need to be expanded in order to enable platform providers to provide scalable yet ethical AI solutions.

Future research will face a unique difficulty in designing responsible AI since it will need us to imbue the system with human and social values in a way that users would recognize and value . On the other hand, a lot of recent research has concentrated on the technical applications of FATE. For instance, a lot of explainable AI research provides technical means of creating explanations. An interpretive process begins when the user is given an explanation. An explanation's interpretability refers to the process by which the user will come to their own independent understanding of it. This interpretation might or might not match the expected meaning that the system's inventor intended.

Consequently, further investigation is required to have a deeper comprehension of the relationship between distinct design patterns, technical solutions, explainability research, and user interpretability. Some user or job factors, for instance, affect how interpretable a user is; for instance, an expert needs different explanations than a novice does. We propose that data provenance warrants additional attention as well, especially in the XAI community, since it offers significant supplementary facts that are essential for user interpretation. Future studies could provide precise criteria, functionalities, and design tenets for creating AI systems that are responsible.

CONCLUSION

To reduce biases and enhance responsible AI-based systems, data provenance is crucial (see Figure 1). Current procedures see data provenance as a requirement of laws, rules, and directives intended to guarantee data availability, confidentiality, and integrity control and protection. Data provenance is seen as an expense associated with adhering to these regulations. These kinds of actions arise from an organization that isn't committed to creating AI-based systems that are accountable.

On the other hand, our suggested approaches see data provenance as a crucial part of creating AI-based systems that are responsible. Organizational performance is likely to increase over time for those that are strategically dedicated to achieving their FATE goals. Our suggested practices understand that losing data provenance at any point in the provenance chain results in a loss of data provenance in all subsequent sections and see data provenance as an investment required to fulfill their FATE goals. As a result, businesses must understand how crucial it is to establish a thorough provenance for crucial data that is used as an input by AI systems.

Data repurposing is increasingly common in modern systems development initiatives, such as those involving data-driven development and AI engineering. Organizations will gain a great deal from data provenance when recommended practices are followed, since the data provenance created for one project is likely to benefit several projects that use the same data. Because various projects frequently use the same data sources, businesses must adopt a holistic approach when analyzing the costs and advantages of data provenance. While recommended practices acknowledge the necessity to preserve dynamic data provenance information that is updated over the data's lifecycle, existing approaches treat data provenance records as static.

Beyond the FATE characteristics, we have listed the many advantages of data provenance. However, companies will have to rank the importance of their data provenance investments according to factors like the size of the rewards from reaching FATE and the gravity of the drawbacks or failure cost from failing to achieve FATE. When faced with financial or time restrictions, organizations that consider data provenance as an overhead expense are more inclined to overlook it and, worse, might participate in unethical behaviors like virtue washing .

Investments in data provenance ought to be motivated by an internal desire to raise the accountability of AI-powered systems. Adopting data provenance techniques, for instance, is beneficial for achieving transparency since it allows users to comprehend, interact with, and audit the AI-based system and its results. Similarly, by defining accountability and preventing harm from deterrence, data provenance that facilitates accountability serves as a way to guarantee justice [15]. These illustrations demonstrate how important FATE traits are to maintaining the inherent values of fundamental concepts like justice and human autonomy. Additionally, businesses that adopt a lifecycle approach understand that early data collecting and processing expenditures pay off later in the lifecycle of an AI-based system. However, although rapidly outweighing the drawbacks, these advantages—such as enhancing reputation, preventing reputational damage, and developing the intended FATE characteristics—are frequently hard to measure.

REFERENCES

[1] 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

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

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

[4] 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

[5] 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

[6] 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

[7] 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).

[8] 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

[9] 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.

[10] 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

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

[12] 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

[13] 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