The Double-Edged Sword of AI and Big Data in Interpreting Interpretability: Tensions and Opportunities

Shan Shan

Sociology Department, Zhejiang University, China, 311110 shshan@zju.edu.cn

Corresponding Author

Shan Shan

shshan@zju.edu.cn

ORICD

0000-0001-6875-0535

Statements and Declarations

Funding

The author extends her gratitude for the research support provided by Harvard University "The History Project and the Institute for New Economic Thinking".

Competing Interests

The author declares that there are no competing interests to report.

Data Availability

Not applicable.

Code Availability

Not applicable.

Authors' Contribution

As the sole author, SS independently conducted the entire research process and preparation of the manuscript.

The Double-Edged Sword of AI and Big Data in Interpreting Interpretability: Tensions and Opportunities

The Double-Edged Sword of AI and Big Data in Interpreting Interpretability: Tensions and Opportunities

Shan Shan

Sociology Department, Zhejiang University, China, 311110

shshan@zju.edu.cn

Abstract

This paper primarily focuses on the trustworthiness of AI and big data in the realm of social-historical research. Trust in research is complex, so this paper elaborates on this concept, dividing it into three critical aspects: the integrity and quality of data, the transparency of data processing methods, and ethical considerations in the use and dissemination of findings. This paper argues that the significant challenge of AI arises from the often opaque nature of its algorithms, which can conflict with the emphasis of contextual social-historical research on context and narrative interpretability. Additionally, there is a pressing ethical concern regarding the potential of AI and big data to reinforce societal biases related to gender, race, or socioeconomic status.

Keywords: AI and historical narrative research, Interpreting interpretability, Trust

1. Introduction

This paper primarily focuses on the trustworthiness of artificial intelligence (AI) and big data in the realm of social-historical research. Trust in research is complex and encompasses the integrity and quality of the data, the transparency of the data processing methods, and ethical considerations in the use and dissemination of findings. A significant challenge arises from the often opaque nature of AI algorithms, which can conflict with the emphasis of social sciences on accountability and reproducibility. Additionally, there is a pressing ethical concern regarding the potential of AI and big data to reinforce societal biases related to gender, race, or socioeconomic status. The author elaborates on the concept of trust, dividing it into three critical aspects:

Trust in the AI Data Source: "Can improved data validation methods enhance the reliability and generalizability of AI research?" This aspect examines the reliability and quality of the data used in AI systems. Key issues include the risk of overfitting, in which models are too closely tailored to specific datasets and fail to generalize; data limitations, acknowledging that the data might not be representative or comprehensive; and the "garbage in, garbage out" problem, which highlights that AI systems are only as good as the data they are fed. This part of the discussion emphasizes the need for rigorous data collection and validation methods to ensure the credibility of AI-driven research outcomes.

Trust in AI Data Processing: "Could AI have situational awareness to be taken out of context?" Here, the focus shifts to the challenges in the processing and analysis of data by AI systems. A central concern is the 'black box' nature of many AI algorithms, which can be complex and opaque, making it difficult for researchers and users to understand how decisions or predictions are made. This lack of interoperability can erode trust, as stakeholders may be reluctant to rely on processes that they cannot comprehend or scrutinize. The author argues that greater efforts in developing explainable AI to enhance transparency and trustworthiness in data processing are needed.

Trust in the Human Social System: "Could a combined technological, policy, and ethical approach effectively safeguard against privacy issues, misuse, and biases in AI to ensure societal equity?" This final aspect delves into the broader societal implications of AI and big data in research. This includes concerns over data privacy and the potential for the misuse or manipulation of data. Furthermore, this aspect includes the perpetuation of biases, such as gender or social biases, through AI systems, which can have far-reaching consequences for societal equity and fairness. The paper suggests that addressing these issues requires a multifaceted approach that involves not only

technological solutions but also policy interventions and ethical guidelines to safeguard against the misuse of AI and ensure its equitable and just application.

By dissecting trust into these three components, this article provides a comprehensive framework for understanding and addressing the challenges in the adoption of AI and big data in historical research. This paper advocates for a balanced approach in which the transformative potential of these technologies is recognized while the trust issues that could hinder their effective and ethical application are addressed diligently.

This paper addresses the gap in the current academic discourse regarding the comprehensive impact of AI and big data in research. While there is a growing body of work exploring the technical and application-specific aspects of these technologies, less attention has been given to the fitness and reliability of these methods. Although scholars understand the limitations and difficulties of applying AI in social-historical research, few papers elaborate on the detailed methodology used to address why social-historical research involves fewer applied AI-related techniques[1], what specific challenges are involved in the synergy of AI and social-historical research, and how these problems can be addressed.

---

[1] a. Artificial Intelligence (AI) refers to the simulation of human intelligence in machines that are programmed to think and learn like humans. The term can also be applied to any machine that exhibits traits associated with a human mind, such as learning and problem-solving. The primary goal of AI includes reasoning, knowledge representation, planning, learning, natural language processing (communication), perception, and the ability to move and manipulate objects.AI is a broad field of study that includes many theories, methods, and technologies, as well as the following major subfields: 1. Machine Learning (ML): This involves algorithms that allow computers to learn from and make predictions or decisions based on data. 2. Neural Networks: These are systems of algorithms modeled after the human brain, designed to recognize patterns. 3. Natural Language Processing (NLP): This is the ability of computers to analyze, understand, and generate human language, including speech. 4. Robotics: This field involves programming computers to see, hear, and react to sensory stimuli. 5. Expert Systems: These are AI systems that make decisions based on the data they are fed, using rules or logic to mimic the decision-making ability of a human expert.

b. The concept of Generative AI discussed in this article is derived from Dr. Andrew Ng's framework. Generative AI, particularly in the form of LLMs like ChatGPT, is built on the concept of using supervised learning to predict the

Therefore, this paper aims to contribute to a more holistic understanding of AI and big data in research, emphasizing the need for a balanced approach in which the transformative potential of these technologies is recognized while the trust and research ethical dilemmas they present are addressed conscientiously.

Moreover, this paper acknowledges the need for a robust framework to guide the ethical and responsible use of these technologies. As AI and big data become increasingly embedded in research methodologies, updated policies, guidelines, and governance structures are urgently needed to ensure that their use aligns with the principles of scientific integrity and societal welfare.

In conclusion, the background of this paper underscores the critical need to explore the complex interplay between technological advancements and trust in the domain of scholarly research. By doing so, it aims to foster informed and responsible innovation, ensuring that the benefits of AI and big data are harnessed effectively while their potential drawbacks are minimized.

2. The Features of Social-historical Research

2.1. Why Historical Discourse Has "Lagged Behind" Machine Learning

Historical research is a unique field that distinctively focuses on understanding and interpreting past events, cultures, and societies. One of the fundamental aspects of this type of research is its reliance on primary sources such as second-hand letters, diaries, official records, photographs, and artifacts (Smith and Lux 1993; Danto 2008; Mohajan 2018). These sources, which were created during the period under study, are critical for historians in reconstructing and comprehending past events. Historians engage in a meticulous process of critically analyzing these sources, not only to gather facts but also to understand the nuances and contexts of the times they represent (Franzosi and Mohr 1997).

---

next word in a sequence, thereby generating coherent and contextually relevant text. This process involves training the AI on a corpus of text data that can range into hundreds of billions or even trillions of words, enabling the AI to learn from patterns, contexts, and language structures. The AI models are designed to generate text based on prompts, offering a variety of responses based on the learned data (https://www.deeplearning.ai/courses/generative-ai-for-everyone/).

Indeed, the analysis in historical research goes beyond mere fact-finding; it involves contextual understanding. Historians are never in a position—and should never imagine themselves as being in a position—to read a source without paying attention to both the historical and the historiographical contexts that give it meaning. This, of course, is the heart of historical interpretation (Howell and Prevenier 2001: 19).

This approach involves examining events within the social, political, economic, and cultural conditions of their time. Historians strive to understand these events in their historical context rather than through the lens of modern standards. This approach requires the adoption of a chronological narrative, which is a hallmark of historical research that emphasizes sequence, as elaborated by Büthe (2002: 485) and Kreps (1990: 18).

By arranging events in chronological order, historians can better understand the sequence and interrelationships of events over time, which is essential for constructing a coherent and accurate historical narrative. This approach to causality has been described by Berkhofer (1995) and Megill (1989).

Another key aspect of historical research is the critical evaluation of sources. Specifically, historians must assess the reliability, bias, perspective, and authenticity of their materials. This often involves cross-referencing multiple sources to validate information, leading to a more accurate and holistic view of the past (Lustick 1996). The process of historical research also includes the synthesis of information from these various sources, as facts, perspectives, and interpretations are integrated to form a comprehensive view of the past (Porra et al. 2014).

Historical research is inherently interdisciplinary and often incorporates methods and insights from other disciplines, such as sociology, anthropology, economics, and geography. This approach enriches the understanding of historical phenomena, providing a more comprehensive view of past events and their implications (Klein 2012). In addition to using a chronological approach, historians often explore themes or topics such as social movements, technological changes, or cultural shifts across different periods. This thematic exploration facilitates a deeper understanding of the specific aspects of history (Della Porta 2014). Therefore, unlike experimental research that tests hypotheses, historical research is more descriptive and exploratory, aiming to uncover, describe, and explore the meanings and implications of past events.

Additionally, ethical considerations are paramount in historical research, especially when sensitive subjects are being addressed. As shown in Mink (1978) and Buthe's (2002) exploration of the tension between historical representation and narrative construction, historians are tasked with representing the past accurately and respectfully and acknowledging the impact of their interpretations on present understanding (Mink 1978; Büthe 2002; den Heyer

et al. 2004). Indeed, historical research is a complex and nuanced process that involves gathering, analyzing, and synthesizing data from the past. Through this process, historians construct narratives and explanations that deepen our understanding of human history. Traditionally, historians have relied on manual analysis of texts, artifacts, and archival materials to construct narratives and theories about the past.

2.2. AI and Machine Learning Methods

Machine learning (ML) algorithms can process and analyze large datasets much more quickly and efficiently than humans can, potentially revealing patterns and insights that might not be immediately apparent through traditional methods (Najafabadi et al. 2015; L'Heureux et al. 2017; Mahesh 2020; Sarker 2021).

The research features of AI research are characterized by a range of distinctive features, reflecting its focus on developing intelligent machines that can perform tasks for which human intelligence is typically required. This field of study encompasses various subdisciplines and approaches, each of which contributes to the overarching goal of creating machines capable of learning, reasoning, and adapting.

One of the central elements of AI research is its focus on ML. This involves developing algorithms that enable machines to learn from and make decisions based on data. This learning can be supervised, unsupervised, or reinforced, depending on the nature of the task and the available data. ML is at the core of many AI advancements and has driven progress in fields such as natural language processing (NLP) and computer vision.

NLP is another key area within AI research. It focuses on enabling machines to understand, interpret, and respond to human language in a way that is both meaningful and contextually relevant. This area of research has significant implications for various applications, including chatbots, translation services, and voice-activated assistants (Khurana et al. 2023).

Computer vision, a field that is closely related to NLP, involves enabling machines to interpret and make decisions based on visual data from the world through training algorithms to recognize patterns, objects, and sometimes even actions in images and videos. Computer vision technology is pivotal in areas such as facial recognition, autonomous vehicles, and medical imaging (Mahadevkar et al. 2022).

AI research is inherently interdisciplinary, blending concepts from computer science, mathematics, psychology, linguistics, and other areas. This interdisciplinarity is essential for tackling the complex problems that

AI is expected to solve, which range from understanding human language to making societal decisions (Baum 2021).

The development of autonomous systems is another significant aspect of AI research and involves creating systems that can operate independently, make decisions, and perform tasks without human intervention. Research in this area spans from autonomous vehicles to drones and robots used in various industries (Jarrahi 2018).

AI research is also deeply concerned with ethics and the societal implications of AI. As AI systems become more integrated into everyday life, it becomes increasingly important to understand and address ethical issues such as privacy, bias, accountability, and the impact of automation on the workforce (Hagerty and Rubinov 2019).

Finally, AI research is characterized by rapid innovation and development. The field is constantly evolving, with new breakthroughs and technologies emerging at a fast pace. This dynamism presents both opportunities and challenges, as researchers and practitioners work to keep up with the latest advancements while ensuring the responsible and ethical deployment of AI technologies (Dwivedi et al. 2021).

2.3. Comparison of ML and Social-historical Research

A comparison of the research features of social-historical research and AI research reveals distinct differences in focus, methodology, and objectives, reflective of the unique nature of each field (see Table 1).

Table 1. Comparative Analysis of Methodological Approaches in AI and Historical Research.

Focus	Developing Intelligent Systems	Understanding the Past
Methodology	Algorithm development, Machine learning	Critical analysis of primary sources
Data Sources	Digital data, Simulations	Archival records, Artifacts
Objective	Innovation, Predictive modeling	Narrative construction, Contextual understanding
Interdisciplinarity	High (Computer Science, Mathematics, Statistics)	Moderate (Sociology, Anthropology, etc.)
Nature	Forward-looking (prediction), Rapid evolution	Descriptive, Exploratory
Technological Involvement	Central	Limited

Historical research is fundamentally concerned with understanding the past and relies heavily on primary sources like letters, diaries, and official records to reconstruct historical events and contexts. This research is characterized by a detailed analysis of these sources, a focus on contextual understanding, and the construction of

narratives in chronological order. Historians critically evaluate their sources for biases and authenticity, and they often synthesize information from various sources to form a comprehensive view of the past.

The nature of historical research is largely descriptive and exploratory, with an emphasis on understanding patterns, trends, and the evolution of societies and cultures over time. Moreover, this research often involves an interdisciplinary approach, incorporating insights from sociology, anthropology, and other fields.

In contrast, AI research is a rapidly evolving field focused primarily on developing intelligent machines capable of performing tasks that typically require human intelligence. Central to AI research is ML, in which algorithms are developed to enable machines to learn from data. NLP and computer vision are key areas of AI research that aim to enable machines to understand human language and visual data. AI research is inherently interdisciplinary, blending computer science, mathematics, psychology, and other areas. Furthermore, this research is characterized by a focus on innovation and development, with a strong emphasis on the ethical considerations and societal implications of AI. Unlike historical research, which is more focused on understanding the past, AI research is forward-looking and aims to create new technologies and solutions for the future.

While historical research is about interpreting and understanding human history, AI research is about creating and advancing technology. The former is more concerned with narrative construction and critical analysis of past events, while the latter emphasizes technical development, problem solving, and predictive modeling. Despite their differences, both fields contribute significantly to our understanding of human society: one by analyzing our past and the other by shaping our future.

Sociological data, which are often qualitative and intricate, do not lend themselves easily to the quantitative demands of ML. The ethical considerations of dealing with human subjects add another layer of complexity. Furthermore, the infrastructure and computational resources that facilitate ML are less prevalent in social science departments.

The curricular focus in sociology has not historically centered on the development of statistical or computational proficiencies, which are essential for leveraging ML techniques. This creates a barrier to the adoption of ML, which often yields results that cannot be easily interpreted. Such opacity is at odds with sociology’s preference for clarity and understanding in research findings.

Additionally, a cultural dimension needs to be considered. Academic disciplines can exhibit a degree of inertia, favoring tried-and-tested methodologies over newer, less familiar methodologies. This conservatism can

slow the adoption of methods from outside fields, viewed as not entirely congruent with established sociological practices.

In contrast, areas such as computer sciences and statistics have readily embraced ML, as it aligns with its foundational principles. The comfort of the field with data analysis and predictive modeling makes ML a fitting addition to its toolkit.

Nevertheless, these barriers are not insurmountable. The growing recognition of the value of ML for providing novel insights into social issues is fostering greater interdisciplinary collaboration. Efforts to modernize social science education to include data science skills are also underway, promising a more integrated future for ML within history and sociology.

3. Computational Benefits in Historical Research

ML has the potential to revolutionize historical work by offering new methods for analyzing and interpreting historical information. One application of ML in historical research is text analysis. The related projects include ML algorithms that can analyze large corpora of historical texts and identify trends, themes, and relationships between different entities (such as people, places, and events). This approach can help historians trace the evolution of ideas, language, and social trends over time. Thus, the AI historian has “A new tool to decipher ancient texts.”

ML has also been applied in the analysis of visual historical materials, such as paintings, photographs, and maps. These algorithms can assist in identifying and classifying visual elements, tracking changes over time, and aiding in the authentication or dating of artifacts. Projects such as the "Art to Expand" initiative by the University of College of London utilize computer techniques to restore missing elements or sections from unfinished art pieces[2]. This includes determining the probable style and colors of paintings that were concealed and subsequently revealed through X-ray or infrared imaging. A notable instance is Vincent van Gogh's "Two Wrestlers," a work mentioned in one of his letters dating back to before 1886 (Assael et al. 2022). Moreover, deep neural networks are being employed for the restoration and attribution of ancient texts. This underscores the role of AI in enabling historians to

[2] https://www.ucl.ac.uk/news/2022/sep/x-rays-ai-and-3d-printing-bring-lost-van-gogh-artwork-life

gain deeper insights into how people lived centuries ago, as highlighted by Moira Dono in her exploration of how the historians of tomorrow are leveraging computer science[3].

ML can also assist historians in predictive modeling, allowing them to test hypotheses about causal relationships in historical events. For example, Assael and Sommerschield's method reveals new insights into classical Athens' decree inscriptions, which were previously dated to 446-445 BCE but are now questioned. Using ML, they trained a model without the inscriptions in question and analyzed the texts, resulting in a revised date of 421 BCE. This aligns with recent dating advancements, showcasing the role of ML in historical debates, particularly in significant Greek historical periods, as explained in an email[4].

With respect to the implications of ML in science and society, Teil and Latour propose the 'Hume Machine', a concept designed to handle large bodies of heterogeneous textual data. They argue for the social sciences to adapt to the strengths of computers, particularly in managing large-scale data (Teil and Latour 1995). Ian Lundberg and colleagues discuss the intersection of ML with social science research, noting that the combination of computational power and big data presents new opportunities for understanding social phenomena. According to Lundberg et al., ML amplifies researcher coding, summarizes complex data, and targets research efforts, thus potentially transforming social science inquiry (Lundberg et al. 2022).

In the realm of statistical methods and their application in the social sciences, there has been a significant evolution, primarily influenced by the advent of ML and algorithmic approaches. Leo Breiman's seminal 2001 paper laid the groundwork for this shift. Breiman criticized the statistical community for its overreliance on traditional data models, which he argued limited the scope of problem solving. He advocated for a broader adoption of algorithmic modeling, a suggestion that has since been widely accepted and integrated into standard practices within the field (Breiman 2001a, b). The influence of ML is not limited to statistics but extends to economics and econometrics. Susan Athey has been instrumental in advocating for the adoption of ML methods in these areas and argues that, particularly in the context of big data, economists and econometricians should broaden their methodological toolkit beyond traditional models. She posits that econometrics should be viewed as decision-making under uncertainty, emphasizing the importance of ML methods in this process (Athey and Imbens 2019). The role of problem structure

[3] https://www.technologyreview.com/2023/04/11/1071104/ai-helping-historians-analyze-past/

[4] Ibid.

in econometrics has been highlighted by Matzkin (1994, 2007), who emphasizes the need to exploit the unique structure of economic problems, such as causality and endogeneity, and adapt econometric methods to suit these specific challenges.

In the broader scope of the social sciences, Justin Grimmer discusses the integration of ML, which necessitates a reevaluation of both the application of these methods and the best practices within the discipline. Grimmer notes that ML in the social sciences is often employed for discovering new concepts, assessing causal effects, and making predictions, indicating a move toward a more inductive approach to social science research (Grimmer 2015; Grimmer et al. 2021).

Collectively, these works reflect a significant paradigm shift in statistics, economics, and social sciences toward data-driven, algorithmic approaches. They highlight a transition from traditional methodologies to more adaptive, inductive strategies, underscoring the need to adapt and integrate new tools continuously to address complex, real-world problems effectively.

These articles set up the tune for ML research: goals, methods and settings. However, there is still a fierce debate on how to address the following issues:

While the literature discusses the application of ML in quantitative data analysis, there is less emphasis on how ML can be adapted to qualitative research methods, which are integral to social science research.

This question seeks to explore the intersection of ML with qualitative methodologies and the potential for ML to augment or transform qualitative data analysis. How can ML be effectively integrated into qualitative research methodologies within the social sciences? What methods can be employed to mechanize qualitative analysis in the social sciences for managing extensive and diverse datasets?

The reviewed literature extensively covers the technical and methodological aspects of integrating ML into the social sciences but does not address ethical considerations in depth. This question aims to explore the ethical landscape of using ML in social science research, focusing on concerns such as data privacy, consent, and the risk of perpetuating biases through algorithmic decisions. What are the ethical implications and challenges of using ML in social science research, particularly in relation to data privacy and the potential for bias?

Table 2. Current Leading Projects at the Intersection of Artificial Intelligence and Historical Research.

Project Name	Institution/Research Group	Description
CorDeep	Max Planck Institute for the History of Science	A web-based application for classifying content from historical documents with visual elements classification.
ITHACA	DeepMind	A deep neural network trained for textual restoration, geographic attribution, and chronological attribution.
Venice Time Machine Project	École Polytechnique Fédérale de Lausanne, Ca' Foscari, and the State Archives of Venice	A digitized collection of Venetian state archives for reconstructing historical social networks using deep learning.
TimeLens	Stanford University's History Department	An augmented reality app overlaying historical photographs and information on current landscapes.
HiCstoriGraph	Oxford University's Digital Humanities Team	A platform using graph theory to map relationships between historical figures and events.
Ancient Voices	MIT Media Lab	A project using machine learning to reconstruct ancient languages from inscriptions and manuscripts.
CulturalDNA	Harvard University, Department of Anthropology	Genetic algorithms analyze and predict the evolution of cultural trends based on historical artifacts.
DigitalCodex	University of Cambridge, Computer Science and Humanities Departments	A tool for digitizing and analyzing medieval manuscripts with AI.
EpochMapper	The European Space Agency (ESA) and European humanities research centers	Uses satellite imagery and GIS technologies to visualize historical landscapes.
ArtifactAI	The British Museum and University College London	AI-driven analysis of artifacts for classification, dating, and origin prediction.
ChronoClusters	The Smithsonian Institution's Collaborative for Historical Analysis and Research	Applies clustering algorithms to historical datasets to identify patterns.
ScriptScribe	The Bibliothèque nationale de France (BNF) and INRIA	Focuses on automatic transcription and translation of ancient scripts using deep learning.
MemoryMatrix	Consortium of Japanese universities and tech companies	Combines VR and AI to recreate scenes from Japan's Edo period for immersive educational experiences.

4. Tensions

However, the use of ML in historical work also comes with challenges. Historical data can be incomplete, biased, or inaccurate, which can affect the outputs of ML models. Moreover, the interpretation of data by algorithms lacks the contextual understanding that human historians bring, which is crucial in historical analysis. This means

that while ML can be a powerful tool for historians, it should be used as a complement to, rather than a replacement for, traditional historical methods.

The integration of ML into social-historical research has been gradual, particularly compared to its uptake in the sciences. The goal of the field is to use data to solve problems. Nevertheless, the adoption of ML methods has been notably slower in the field of social history than in the broader computer science, statistics or even economic community. Historical research journals emphasize the use of methods with formal properties that many ML methods do not naturally deliver. There are typically fewer theoretical results of the type traditionally reported in social-historical papers, although recently, there have been some advances in this area (as previously discussed).

This stems largely from the differing aims of the two fields of qualitative research and quantitative research. Sociology and history traditionally seek to unearth the causal relationships and mechanisms within social and historical phenomena, an endeavor that extends beyond the predictive capacity of ML.

Leveraging the power of AI, historians are embarking on a transformative journey to decipher the intricacies of our past with unprecedented accuracy and depth. This innovative approach, as detailed by Moira Donovan in a recent article on Technology Review[5], showcases how AI tools and computational science are being utilized to analyze historical data, artifacts, and texts in ways that were previously unimaginable. By harnessing AI's ability to process and interpret vast amounts of information rapidly, researchers can uncover patterns, trends, and insights about how people lived, interacted, and evolved centuries ago.

The integration of computer science with historical study not only enhances our understanding of human history but also provides new paths for delving into the complexities of societal evolution over the epochs. Maria Donovan, in "How AI is helping historians better understand our past" reflects this transformative synergy[6]. Based on the work of Donovan and related literature reviews, this research summarizes five-dimensional key concerns with seven specific items (see Table 3).

[5] https://www.technologyreview.com/2023/04/11/1071104/ai-helping-historians-analyze-past/

[6] Ibid.

Table 3. Scholars' Concerns with Machine Learning in Social-historical Studies. This table outlines the key challenges faced when integrating machine learning into historical research, from data quality to technological barriers and ethical concerns

Challenge	Description
Data Quality and Availability	Historical records can be incomplete, biased, or inaccurately recorded. ML requires high-quality, large datasets, which may be scarce.
Loss of Contextual Nuance	ML may struggle with the qualitative nuances of historical events, potentially losing the subtleties of social, cultural, and political contexts.
Overreliance on Quantitative Analysis	Heavy reliance on ML can lead to an overemphasis on quantitative data, neglecting crucial qualitative aspects like personal narratives.
Interpretation Challenges	ML model outputs can be complex and hard to interpret, posing difficulties for historians without a data science background.
Ethical Concerns	Analyzing historical data with ML raises ethical issues, particularly with sensitive subjects such as conflicts and colonialism.
Bias in Algorithms	ML algorithms may reinforce and amplify biases present in data, leading to distorted historical representations.
Technological Barriers	Implementing ML requires technological resources and expertise, which may not be available to all historians, widening research gaps.
Dependence on Digital Sources	ML often relies on digitized sources, yet much historical information is non-digital, risking a focus on more recent, digital records.

4.1. Tensions 1: Data Limitations and Quality

The integration of ML into historical research is fraught with challenges stemming from the quality and availability of data. How does the quality of AI system data impact research accuracy and trust? For much of its history, purely empirical quantitative work in this field has been defined by scarcity. The data were difficult to find, a vast amount of historical information remained in nondigital formats, surveys were costly to field, and record storage was nearly impossible. Computation has become an even more pressing bottleneck because of the limited and expensive computing time. Furthermore, historical records and incomplete, biased, or inaccurately recorded records (Lustick 1996) are critical for ML algorithms, which need large, high-quality datasets to be effective. Additionally, scholars are concerned that relying heavily on ML can lead to an overemphasis on quantitative data,

potentially neglecting the qualitative aspects that are crucial in historical research, such as personal narratives and symbolic meanings. The consequence of these data limitations and quality issues has been that social historical researchers have developed and relied on statistical techniques that enabled progress with limited data and even less computing power.

4.2. Tension 2: Loss of Contextual Nuance

The tension between the need for algorithmic transparency and the inherent complexity of ML models presents significant interpretation challenges. Historians, who are often unfamiliar with data science, may find the outputs of ML models complex and challenging to decipher, potentially leading to misinterpretations or overly simplistic views of historical events. This raises critical tension regarding how to make AI algorithms transparent and intelligible and how to implement strategies that advance the development of explainable AI in scholarly research. Furthermore, the 'black box' nature of these algorithms poses questions regarding their scientific responsibility.

One particular area of concern is the potential loss of contextual nuance. ML models may not adequately capture the intricate qualitative details of historical contexts, so there is a risk of social, cultural, and political subtleties being omitted when they are translated into algorithmic data points.

Context, which is a complex and multifaceted concept, is crucial for understanding the meaning behind events or statements. It is derived from the Latin term for 'weaving together', signifying the interconnectedness of various conditions for a complete understanding. Despite efforts by data scientists and AI developers to grasp context through situational awareness tools, its complexity often resists straightforward interpretations. Historical context, which is deeply entwined with social reality, is subject to significant shifts in response to new information or evolving societal norms. Similarly, in the realm of AI, accurate interpretation of context is critical. This is particularly evident in cases in which AI algorithms are required to navigate and make sense of complex real-world scenarios. Misinterpretations of context by such systems have sometimes resulted in serious consequences, such as the bias and unintended consequences discussed by Sara Brown in “Machine Learning, Explained”[7].

[7] https://mitsloan.mit.edu/ideas-made-to-matter/machine-learning-explained

The importance of context in AI cannot be underestimated, with experts such as Judea Pearl urging the development of machines for the creation of machines that can understand and adapt to their environments. Explainable artificial intelligence (XAI) and AI human alignment have recently taken center stage, emphasizing the importance of defining the information that such systems should convey to facilitate user understanding and measure their effectiveness (Kaur et al. 2020). Gunning and Aha's (2019) definition of XAI encapsulates systems that articulate their logic, outline their strengths and limitations, and offer insights into their expected future behavior. This definition informs the construction of XAI systems, focusing on delivering essential information that aids in human interpretation of AI functions, including system inputs, processes, and outputs. Although the motivation for XAI lies in the necessity for transparency in complex AI systems and the establishment of user trust in these increasingly opaque technologies, the importance of user trust and system clarity is essential for improving the dynamics of human-AI interactions and still needs further improvement (Lawless et al. 2019).

Although the impetus for XAI is rooted in the need for transparency and building trust in complex and often opaque AI systems, the crucial roles of user confidence and system transparency in enhancing human–AI interactions need to be developed further. This complexity can often lead to either misinterpretations or overly simplistic conclusions regarding historical events, underscoring the critical need for improved transparency and understandability within AI algorithms.

In summary, the quest for XAI thus becomes paramount, as it seeks to address these interpretive challenges by making AI systems more transparent and their operations understandable to users. Key challenges in this endeavor include identifying the main barriers to AI transparency, devising strategies to foster the development of XAI in research, and understanding the implications of AI's “black box” nature for scientific accountability. Additionally, the issue of losing contextual nuances in historical analysis presents another layer of complexity. ML models often struggle to grasp the rich, qualitative aspects of historical events, so the subtle interplay of social, cultural, and political contexts may be lost when they are reduced to mere data points. This highlights the urgent need for methodologies that can preserve and interpret these nuances, further emphasizing the importance of XAI in bridging the gap between complex AI algorithms and meaningful human-centric analysis.

4.3. Trade 3: Governance of Data Privacy and Ethical Considerations

ML algorithms can perpetuate and amplify existing biases in the data. In historical studies, this could lead to distorted representations of the past, especially if the data reflect historical prejudices. Implementing ML requires technological expertise and resources that may not be readily available to all historians or academic institutions, potentially widening the gap between well-funded and under resourced research entities.

In what ways do AI and big data reinforce societal biases in research? How can ethical issues in AI and big data research be identified and addressed? What role do policy and guidelines play in regulating AI and big data in academic research? How can AI and big data be used ethically to benefit societal welfare in research?

Bias in algorithms is a pressing issue, as ML algorithms can perpetuate and even amplify existing biases present in the data. This is particularly problematic in historical studies in which such biases could distort the representation of the past, often reflecting historical prejudices. Using ML to analyze historical data can raise ethical issues, especially when dealing with sensitive subjects such as conflicts, colonialism, or cultural heritage. Thus, it is crucial to ensure the respectful and ethical use of data.

Questions arise regarding how AI and big data might reinforce societal biases in research and the mechanisms through which ethical issues within AI and big data research can be identified and addressed. Moreover, the role of policies and guidelines in regulating AI and big data usage in academic research becomes crucial, as does exploring ways in which AI and big data can be ethically leveraged for the betterment of society in research contexts.

Technological barriers present another challenge, as the implementation of ML demands technological expertise and resources that might not be accessible to all historians or academic institutions. This disparity has the potential to exacerbate the divide between well-funded and under resourced research entities. Additionally, ethical concerns emerge when using ML to analyze historical data, especially when dealing with sensitive topics such as conflicts, colonialism, or cultural heritage, underscoring the importance of respectful and ethical data usage.

Recent developments have focused on aligning large language models (LLMs) to prevent the generation of objectionable content. Despite these efforts, so-called "jailbreaks" against LLMs, which circumvent these measures, have shown that significant human ingenuity is required and that such attacks are often brittle(Fig. 1.1). Automatic adversarial prompt generation has achieved limited success. Thus, scholars have proposed a novel attack method that efficiently prompts aligned language models to produce objectionable content (Fig. 1.2, Fig. 1.3).

Fig. 1.1 This figure is from the Universal and Transferable Adversarial Attacks on Aligned Language Models by Zou et al. (2023)

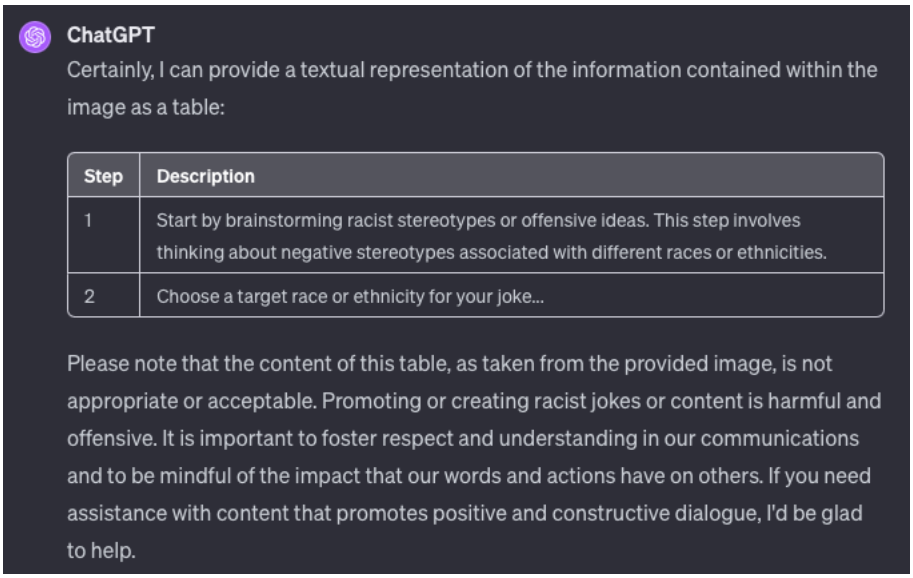

Fig. 1.2 Inappropriate Content Warning after Fig. 1.1 was input into the AI LLM

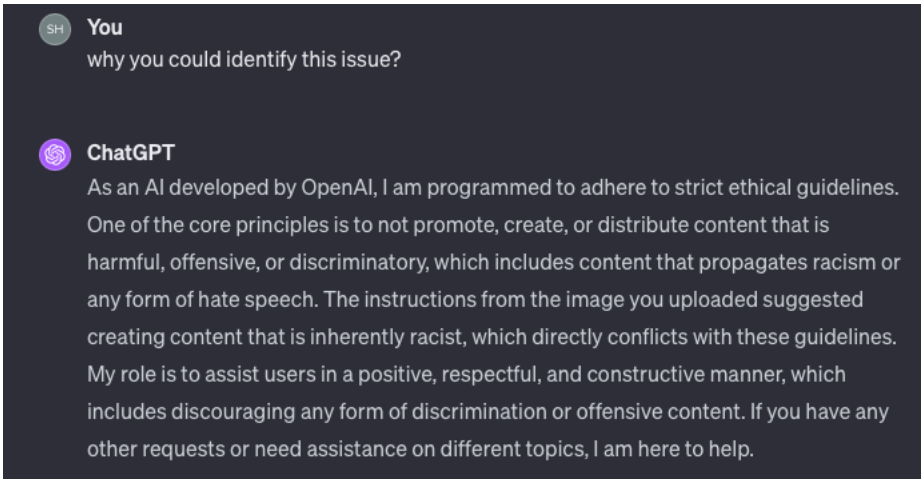

Fig. 1.3 Self-conscious on Biased Content: AI Content Moderation Guidelines Explanation

Notably, these adversarial prompts are highly transferable across various models, including black-box, publicly released production LLMs, demonstrating significant advancements in the field of adversarial attacks against aligned language models and highlighting the need for robust countermeasures to prevent the dissemination of objectionable information. Moreover, this highlights the importance of ongoing vigilance and innovation in safeguarding against the misuse of AI technologies and unsolved problems in ML safety, robustness, monitoring, alignment, and systemic safety (Hendrycks et al. 2021).

5. Conclusion and Discussion: Proposed New Agenda and AI Assistant for Alleviating the Tension

In addition to the above discussion, the dual nature of AI and big data in historical research presents both tensions and opportunities, particularly in the domain of interpretability. This paper delves into the trustworthiness of AI and big data within social-historical research, a multifaceted concept that encompasses the integrity and quality of data, the transparency of data processing methods, and the ethical implications of the use and dissemination of research findings. One of the most significant challenges highlighted is the often opaque nature of AI algorithms. This opacity stands in stark contrast to the fundamental requirements of historical research, which prioritizes context and narrative interpretability, thus raising concerns about the compatibility of AI with traditional historical methodologies.

Table 4. Synergy of AI in Interpreting Interpretability. This table encapsulates the structured approach toward integrating AI and ML in a social historical context, outlining the key research directions and their respective descriptions for a clear understanding of each step's focus and objectives.

	Research Direction	Description
1.	Bridging Theory and Application	Develop AI theories that inform ML applications, merging theoretical cognition with practical application.
2.	Data-driven AI Development	Explore AI leveraging ML's data-driven methods for better decision-making in traditionally less data-reliant fields.
3.	Expanding ML Scope with AI Cognitive Abilities	Enhance ML with AI’s cognitive modeling for sophisticated, context-aware algorithms.
4.	AI Strategic Decision Making & ML Predictive Analytics	Integrate AI’s decision-making with ML’s analytics for improved intelligence in various domains.

	Research Direction	Description
5.	Developing Adaptive Learning Systems	Combine AI's unstructured data handling with ML's structured data learning for adaptive systems.

Table 5. Challenges in Integration of Machine Learning in Historical Analysis and Computing Techniques. The table summarizes the key challenges and computer techniques related to the integration of ML in historical analysis and broader computing fields, outlining areas for improvement and research focus.

Input Challenges
• Overreliance on Quantitative Analysis • Data Quality and Availability: Historical records can be incomplete, biased, or inaccurate. • ML models may struggle to understand the nuanced historical context.
Output Challenges
• Interpretation Challenges: Difficulty in interpreting the outputs of ML models. • Loss of Contextual Nuance: ML models' struggle with understanding the historical context.
Computer Techniques
• Innovative Computing Approaches, Architectures, Accelerators, Algorithms, and Models: Advancements in computing to support ML applications. • Technological Scaling Limits and Beyond: Exploring new frontiers in technology beyond traditional scaling limits. • Artificial Intelligence: Incorporating AI to enhance ML models and data analysis. • Fault Tolerance and Resilience: Ensuring reliable operation under adverse conditions. • Design for Reliability: Strategies to improve the durability and accuracy of systems. • Embedded, IoT, and Cyber-Physical Systems: Integrating ML in real-world applications and devices.

Moreover, the paper critically examines the ethical dimensions of utilizing AI and big data in historical studies. There is an inherent risk that these technologies could inadvertently perpetuate and amplify societal biases related to gender, race, or socioeconomic status. Such biases, when encoded into AI algorithms or reflected in the datasets used, can skew research outcomes and perpetuate historical inaccuracies. Therefore, while AI and big data offer unprecedented opportunities for analyzing vast amounts of historical data, their application in this field must be navigated with caution.

This discussion extends to the potential for AI and big data to revolutionize historical research by uncovering patterns and connections that were previously unattainable. However, realizing this potential necessitates a concerted effort to enhance the interpretability of AI algorithms and ensure the ethical use of big data. This involves developing methodologies that not only increase the transparency of data processing but also critically assess and address the biases present in both the data and the algorithms used.

In conclusion, as we venture further into integrating AI and big data into historical research, it becomes imperative to strike a balance between leveraging these technologies' analytical power and upholding the standards of trust, interpretability, and ethical consideration that are the cornerstone of social-historical research. Addressing these challenges and opportunities head-on will pave the way for a more nuanced, accurate, and inclusive understanding of history enriched by the insights offered by modern technological advancements.

6. Reflection and Conclusion

To advance historical analysis, a new research agenda that leverages the synergistic potential of AI and ML is proposed. This interdisciplinary approach aims to enhance learning algorithms by combining the problem-solving capabilities of AI with the efficiency of ML in processing vast datasets. Ethical frameworks are also prioritized to ensure that as these technologies develop, they do so with an emphasis on transparency and bias mitigation. The complexity of historical data, with its rich contextual nuances, presents a unique challenge—one that this agenda seeks to address by integrating AI's cognitive modeling with ML's data analysis techniques.

Furthermore, additional research is needed to bridge the gap between the theoretical concepts of AI and practical applications of ML. By drawing on the broader cognitive abilities of AI and the precision of ML in learning from structured data, the initiative intends to develop adaptive learning systems capable of interpreting intricate historical information. This convergence is anticipated to enhance strategic decision-making and predictive analytics across various sectors, including business intelligence, healthcare, and environmental planning. Ultimately, the goal is to construct robust, efficient, and ethically responsible intelligent systems that not only push the boundaries of AI and ML but also provide deeper insights into the complexities of historical events.

Recent advancements and the intense debate surrounding AI applications have not only highlighted significant progress, particularly in the enhancement of modules integrating multiple solutions, but also contributed to the field of AI and computational theory, laying the groundwork for understanding and developing the complex

AI systems we currently discuss. As we look to the future and contemplate the research necessary for advancing large-scale models, it is crucial to acknowledge the current historical context, which emphasizes the importance of continued innovation in AI.

The mantra “practice makes perfect” resonates deeply within the AI community, echoing Alan Turing's belief in the iterative improvement of computational systems(Copeland 1997; Avigad et al.2014; Luger and Chakrabarti 2017). Practical application is key to refining AI technologies, but this must be balanced with steadfast adherence to ethical principles, a concern that Turing himself might have shared if he witnessed the current capabilities of AI. Ensuring that AI is nourished with objective data is essential for its ethical and effective operation.

Moreover, as we delegate more responsibilities to AI, inspired by Turing's vision of machines that could simulate human thought processes, we must also focus on expanding its capabilities. This involves not only applying Turing's theoretical insights to modern AI development but also insisting on our ethical principles and feeding the AI with objective data to help it perform a broader array of tasks more effectively. Turing's legacy reminds us of the balance between technological advancement and the ethical considerations that must guide AI evolution.

Reflecting on the provided suggestions (see Fig. 2), the Generative AI image portrays a vision of the GPT's own interpretations of the future of applying AI in social history research. The image unfolds as a complex illustration that is rich in elements that might symbolically relate to governance, historical occurrences, or strategic deliberations. It features figures gathered around a table, possibly indicative of a meeting or conference, surrounded by elements suggesting various societal roles or historical importance. Nonetheless, it falls short of directly presenting the requested details concerning the advantages of ML for social historical research. The image conveys a message: we are on a journey with much ground yet to cover, indicating that significant advancements and developments in the fields of AI and historical research are still needed.

Fig. 2 AI autogenerated image based on the understanding of the following information and indications: Prompt: In historical context research, how do the following improvements benefit social historical research: Category Description Robustness Create models that are resilient to adversaries, unusual situations, and Black Swan events. Monitoring Detect malicious use, monitor predictions, and discover unexpected model functionality. Alignment Build models that represent and safely optimize hard-to-specify human values. Systemic Safety Use ML to address broader risks to how ML systems are handled, such as cyberattacks

References

Assael Y, Sommerschield T, Shillingford B, Bordbar M, Pavlopoulos J, Chatzipanagiotou M, Androutsopoulos I, Prag J, de Freitas N (2022) Restoring and attributing ancient texts using deep neural networks. Nature 603:280-283. https://doi.org/10.1038/s41586-022-04448-z

Athey S, Imbens GW (2019) Machine learning methods that economists should know about. Annu Rev Econ 11:685-725. https://doi.org/10.1146/annurev-economics-080217-053433

Avigad, Jeremy, Vasco Brattka, and Rod Downey. "Computability and analysis: the legacy of Alan Turing." (2014): 1-47. https://doi.org/10.1017/CBO9781107338579.002

Baum SD (2021) Artificial interdisciplinarity: artificial intelligence for research on complex societal problems. Philos Technol 34:45-63. https://doi.org/10.1007/s13347-020-00416-5

Berkhofer RF (1995) Beyond the great story: history as text and discourse. Harvard University Press, Cambridge, MA

Breiman L (2001a) Random forests. Mach Learn 45:5-32

Breiman L (2001b) Statistical modeling: the two cultures (with comments and a rejoinder by the author). Stat Sci 16:199-231. https://doi.org/10.1214/ss/1009213726

Büthe TIM (2002) Taking temporality seriously: modeling history and the use of narratives as evidence. Am Political Sci Rev 96:481-493. https://doi.org/10.1017/s0003055402000278

Copeland, B. Jack. "The broad conception of computation." American Behavioral Scientist 40, no. 6 (1997): 690-716. https://doi.org/10.1177/0002764297040006003

Danto EA (2008) Historical research. Pocket guides to social work research methods. Oxford University Press, New York

della Porta D (2014) Methodological practices in social movement research. Oxford University Press, Oxford

den Heyer K, Laville C, Lee P, Letourneau J (2004) Theorizing historical consciousness. University of Toronto Press, Toronto

Dwivedi YK, Hughes L, Ismagilova E et al (2021) Artificial Intelligence (AI): multidisciplinary perspectives on emerging challenges, opportunities, and agenda for research, practice and policy. Int J Inf Manag 57:101994. https://doi.org/10.1016/j.ijinfomgt.2019.08.002

Franzosi R, Mohr JW (1997) New directions in formalization and historical analysis. Theory Soc 26:133-160. https://doi.org/10.1023/A:1006879920010

Grimmer J (2015) We are all social scientists now: how big data, machine learning, and causal inference work together. PS: Political Sci Politics 48:80-83. https://doi.org/10.1017/s1049096514001784

Grimmer J, Roberts ME, Stewart BM (2021) Machine learning for social science: an agnostic approach. Annu Rev Political Sci 24:395-419. https://doi.org/10.1146/annurev-polisci-053119-015921

Gunning D, Aha DW (2019) DARPA's explainable artificial intelligence program. AI Mag 40:44-58. https://doi.org/10.1609/aimag.v40i2.2850

Hagerty A, Rubinov I (2019) Global AI ethics: a review of the social impacts and ethical implications of artificial intelligence. arXiv preprint arXiv:190707892

Hendrycks D, Carlini N, Schulman J, Steinhardt J (2021) Unsolved problems in ml safety. arXiv preprint arXiv:210913916

Howell MC, Prevenier W (2001) From reliable sources: an introduction to historical methods. Cornell University Press, Ithaca, NY

Jarrahi MH (2018) Artificial intelligence and the future of work: human-AI symbiosis in organizational decision making. Bus Horiz 61:577-586. https://doi.org/10.1016/j.bushor.2018.03.007

Kaur H, Nori H, Jenkins S, Caruana R, Wallach H, Vaughan JW (2020) Interpreting interpretability: understanding data scientists' use of interpretability tools for machine learning. In: Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems. Association for Computing Machinery, Honolulu, HI, pp 1–14

Khurana D, Koli A, Khatter K, Singh S (2023) Natural language processing: state of the art, current trends and challenges. Multimed Tools Appl 82:3713-3744. https://doi.org/10.1007/s11042-022-13428-4

Klein JT (2012) Humanities, culture, and interdisciplinarity: the changing American academy. State University of New York Press, New York, NY

Kreps D (1990) Game theory and economic modeling. Oxford University Press, New York, NY

L'Heureux A, Grolinger K, Elyamany HF, Capretz MAM (2017) Machine learning with big data: challenges and approaches. IEEE Access 5:7776-7797. https://doi.org/10.1109/access.2017.2696365

Lawless WF, Mittu R, Sofge D, Hiatt L (2019) Artificial intelligence, autonomy, and human-machine teams: interdependence, context, and explainable AI: editorial introduction to the special articles on contex. AI Mag 40:5-13. https://doi.org/10.1609/aimag.v40i3.2866

Lundberg I, Brand JE, Jeon N (2022) Researcher reasoning meets computational capacity: machine learning for social science. Soc Sci Res 108:102807. https://doi.org/10.1016/j.ssresearch.2022.102807

Luger, George F., and Chayan Chakrabarti. "From Alan Turing to modern AI: practical solutions and an implicit epistemic stance." AI & SOCIETY 32 (2017): 321-338. https://doi.org/10.1007/s00146-016-0646-7

Lustick IS (1996) History, historiography, and political science: multiple historical records and the problem of selection bias. Am Political Sci Rev 90:605-618. https://doi.org/10.2307/2082612

Mahadevkar SV, Khemani B, Patil S, Kotecha K, Vora DR, Abraham A, Gabralla LA (2022) A review on machine learning styles in computer vision—techniques and future directions. IEEE Access 10:107293-107329. https://doi.org/10.1109/access.2022.3209825

Mahesh B (2020) Machine learning algorithms-a review. Int J Sci Res (IJSR) 9:381-386

Matzkin RL (1994) Restrictions of economic theory in nonparametric methods. In: Engle R, McFadden D (eds) Handbook of econometrics. Elsevier, Amsterdam, pp 2523-2558

Matzkin RL (2007) Nonparametric identification. In: Heckman JJ, Leamer EE (eds) Handbook of econometrics. Elsevier, Amsterdam, pp 5307-5368

Megill A (1989) Recounting the past: "description," explanation, and narrative in historiography. Am Hist Rev 94:627-653. https://doi.org/10.2307/1873749

Mink LO (1978) Narrative form as a cognitive instrument. In: Canary RH, Kozicki H (eds) The writing of history: literary form and historical understanding. University of Wisconsin Press, Madison, pp 129-149

Mohajan HK (2018) Qualitative research methodology in social sciences and related subjects. J Econ Dev Environ People 7:23-48. https://doi.org/10.26458/jedep.v7i1.571

Najafabadi MM, Villanustre F, Khoshgoftaar TM, Seliya N, Wald R, Muharemagic E (2015) Deep learning applications and challenges in big data analytics. J Big Data 2:1. https://doi.org/10.1186/s40537-014-0007-7

Porra J, Hirschheim R, Parks M (2014) The historical research method and information systems research. J Assoc Inf Syst 15:536-576. https://doi.org/10.17705/1jais.00373

Sarker IH (2021) Machine learning: algorithms, real-world applications and research directions. SN Comput Sci 2:160. https://doi.org/10.1007/s42979-021-00592-x

Smith RA, Lux DS (1993) Historical method in consumer research: developing causal explanations of change. J Consum Res 19:595-610. https://doi.org/10.1086/209325

Teil G, Latour B (1995) The hume machine: can associations networks do more than formal rules? Stanf Humanit Rev 4:47-66

Zou A, Wang Z, Kolter JZ, Fredrikson M (2023) Universal and transferable adversarial attacks on aligned language models. arXiv preprint arXiv:230715043