A comprehensive review of techniques for documenting artificial intelligence

Florian Königstorfer

Florian Königstorfer is based at the Business Analytics and Data Science Center (BANDAS Center), Karl Franzens University Graz, Graz, Austria.

Abstract

Purpose – *Companies are increasingly benefiting from artificial intelligence (AI) applications in various domains, but also facing its negative impacts. The challenge lies in the lack of clear governance mechanisms for AI. While documentation is a key governance tool, standard software engineering practices are inadequate for AI. Practitioners are unsure about how to document AI, raising questions about the effectiveness of current documentation guidelines. This review examines whether AI documentation guidelines meet regulatory and industry needs for AI applications and suggests directions for future research.*

Design/methodology/approach – *A structured literature review was conducted. In total, 38 papers from top journals and conferences in the fields of medicine and information systems as well as journals focused on fair, accountable and transparent AI were reviewed.*

Findings – *This literature review contributes to the literature by investigating the extent to which current documentation guidelines can meet the documentation requirements for AI applications from regulatory bodies and industry practitioners and by presenting avenues for future research. This paper finds contemporary documentation guidelines inadequate in meeting regulators' and professionals' expectations. This paper concludes with three recommended avenues for future research.*

Originality/value – *This paper benefits from the insights from comprehensive and up-to-date sources on the documentation of AI applications.*

Keywords *Artificial intelligence, AI documentation, AI development, AI governance*

Paper type *Literature review*

1. Introduction

In recent years, artificial intelligence (AI) has had a notable impact on corporations and society. Currently, AI is not only increasing revenue and efficiency (Alfaro *et al.*, 2019) but also assisting in the realm of justice (Dressel and Farid, 2018), human resources (HR) (Kupfer *et al.*, 2023), and numerous other domains. However, AI's downsides have recently become public. For instance, AI biases may lead to discrimination against minority groups (Heinrichs, 2022) and have far-reaching implications for society (Makridakis, 2017). Challenges relating to addressing these issues through governance lead to barriers in the adoption of AI in practice.

Effective information technology (IT) governance relies on transparency (Winter and Davidson, 2019). In software engineering, transparency is promoted through documentation (ISO, 2019; Simonsson *et al.*, 2010). This also applies to AI. For AI governance, documentation can reduce errors, making AI documentation a vital governance tool (Collins *et al.*, 2015; Kapoor and Narayanan, 2022).

However, despite advancements in explainable AI (XAI) (Gashi *et al.*, 2022), there remains a significant gap in research concerning AI documentation for effective supervision. Notably, existing guidelines from software engineering principles do not adequately cater to the needs of AI auditors (Appelbaum *et al.*, 2017) and despite the presence of clear

Received 12 January 2024
Revised 1 April 2024
Accepted 5 April 2024

requirements for AI documentation (European Parliament, 2024; Königstorfer and Thalmann, 2021), studies find that industry experts are uncertain about how to comply with these requirements, viewing it as a major impediment to adopting AI (Königstorfer and Thalmann, 2021, 2022). This shows that there is a pressing need for more comprehensive AI documentation that serves practitioners (European Parliament, 2024; Königstorfer and Thalmann, 2021). Consequently, there remains a risk that AI applications lack sufficient documentation, potentially leading to insufficient governance. As a result, it is crucial to assess whether the current methodologies and instruments for AI documentation are adequate to meet the requirements for AI documentation. To shed light on the capabilities, limitations and future research directions of current AI documentation methods, a structured literature review according to Webster and Watson (2002) will be presented in this paper. The research question is:

RQ1. What methods are available for documenting AI applications? What are the limitations and challenges associated with these methods?

2. Background

In the Past years, AI, especially its powerful subfield machine learning (ML), has seen increased predictive power due to more data and cheaper processing (Sun *et al.*, 2017), aiding in the justice system (Dressel and Farid, 2018), HR (Kupfer *et al.*, 2023) and many other fields. However, AI's negative aspects have become apparent. Biases in AI result in discrimination against minorities (Heinrichs, 2022) and extensive changes to society as a whole (Makridakis, 2017). In addition, many academic AI models, including peer-reviewed ones, have flaws, often hidden because of a lack of transparency (Gundersen and Kjensmo, 2018; Kapoor and Narayanan, 2022).

This issue raises concerns for society (Sadek *et al.*, 2024). The EU's AI Act reflects these societal concerns, proposing a risk-based approach to regulation where certain risky AI applications are prohibited or strongly regulated (European Parliament, 2024). AI applications with potential for surveillance, exploitation or manipulation, like social scoring, subliminal manipulation, indiscriminate biometric identification in public spaces and exploiting vulnerabilities, are banned due to their societal threats. Systems enabling (potentially erroneous) denial of social and medical services, justice and employment, such as biometric identification for social benefits, employment or criminal risk assessments, are allowed under stringent conditions emphasizing documentation, data integrity, transparency and human oversight. Thus, transparent and accountable AI design and robust governance are crucial to align AI with business objectives, legal mandates and societal norms (Ndlovu and Kyobe, 2016; Sadek *et al.*, 2024). However, literature on governance for AI is scarce (Winter and Davidson, 2019). Literature suggest that a suitable documentation is crucial for the ethical, transparent and responsible deployment of AI (Sadek *et al.*, 2024).

Simultaneously, creating AI governance and AI-specific documentation guidelines is challenging due to differences from traditional software, especially in ML. ML models derive decision-making from training data with minimal developer instructions, leading to dependency on data quality and preparation (Gebru *et al.*, 2021). While this enhances predictive capabilities, it also introduces biases and unpredictability (Ellul *et al.*, 2021). Also, documenting ML's decision-making is difficult due to its "Black Box" nature, particularly in advanced models (Gashi *et al.*, 2022). This opacity hinders effective AI governance and real-world integration (Königstorfer and Thalmann, 2021). Hence, traditional software documentation methods, like source code, are less relevant for ML (Garousi *et al.*, 2015). To enhance the readability of this paper, "AI" refers to models with these characteristics and challenges.

In software engineering, documentation records architectural decisions for stakeholders (Clements *et al.*, 2011). This contrasts with XAI, which aims to explain AI's decisions (Gashi *et al.*, 2022). Effective documentation mitigates errors in AI creation and represents a potential governance tool by capturing the decisions made during the development and the deployment of the AI (Collins *et al.*, 2015; Winter and Davidson, 2019). Inadequate documentation results in transparency gaps, impeding accountability and ethical AI usage, because opacity complicates identifying harm and attributing responsibility (Wachter and Mittelstadt, 2019). Consequently, regulators and practitioners place high demands on AI documentation. First, because training data significantly influences AI models (Gebru *et al.*, 2021) and errors often occur during data preparation (Kapoor and Narayanan, 2022), documenting the training data and the data preparation processes is crucial for effective AI governance. Second, AI documentation should record and explain development and design phase decisions, including ML algorithms, feature engineering and model parameters (European Parliament, 2024). This is essential due to the significant errors that can arise from incorrect design decisions. Third, AI documentation should describe the application domain, safety and security systems (European Parliament, 2024; Königstorfer and Thalmann, 2021). It should also detail intended use cases, business process integration and AI's interaction with hardware and other software (European Parliament, 2024; Königstorfer and Thalmann, 2021). Documenting these aspects is vital for ensuring user protection against potential AI errors. However, despite clear requirements for AI documentation, industry professionals are uncertain about meeting them, seeing this as a major barrier to AI adoption (Königstorfer and Thalmann, 2021, 2022). Therefore, it is crucial to verify if current methods for AI documentation can adequately meet these requirements.

3. Method and procedure

To clarify how well existing AI documentation methods and tools can be used to enable the governance of AI applications, a structured literature review, according to Webster and Watson (2002) was conducted. The structured literature review consists of three steps:

1. identification of relevant literature;
2. structuring the review; and
3. theoretical development.

To identify suitable literature, a Scopus query targeted 22 key conferences and journals, including from the AIS Basket of Eight, major information systems conferences (European Conference on Information Systems, Hawaii International Conference on System Sciences and International Conference on Information Systems), medical journals (e.g. *Nature*) and fair, accountable and transparent AI venues (e.g. Association for Computing Machinery Conference on Fairness, Accountability and Transparency), for literature from January 2011 to December 2023. The focus was on IS and medical journals because they document business applications and governance of AI systems better than CS publications and similarly detail AI's technical aspects like training data (Königstorfer *et al.*, 2024).

Keywords for the Scopus query were chosen using Rowley and Slack's (2004) building block approach, focusing on 12 combinations – "documentation," "reporting guideline," "reviewability," "reproducibility," "accountability" and "transparency" each paired with "artificial intelligence" or "machine learning." These terms are significant in AI research and specific fields, like medical research where "reporting guidelines" detail AI solutions. Other keywords like "safety" and "artificial intelligence" were tested but found irrelevant to AI documentation (Rowley and Slack, 2004).

The Scopus query yielded 382 potential publications, and additional Google Scholar queries found 173 more. Following Webster and Watson (2002), an abstract scan was conducted. Specifically, all papers containing concrete methods for meeting at least one of the requirements presented in the previous chapter were included. Papers not in English or inaccessible were excluded. This process selected 31 papers and a forward-backward scan added seven more, totaling 38 reviewed papers.

A concept-centric approach structured the review, using Webster and Watson's (2002) concept matrix. This matrix focuses analysis on relevant concepts rather than authors. A qualitative content analysis, per Patton (2014), extracted data patterns, refined through repeated analysis. Table 1 presents the review dimensions (Patton, 2014).

4. Results

The AI documentation requirements form the basis of this section. Each subsection outlines methods to document requirements and paper limitations. Figure 1 illustrates the section's structure.

4.1 Documenting the training data

The literature offers various tools and methods for documenting the training data.

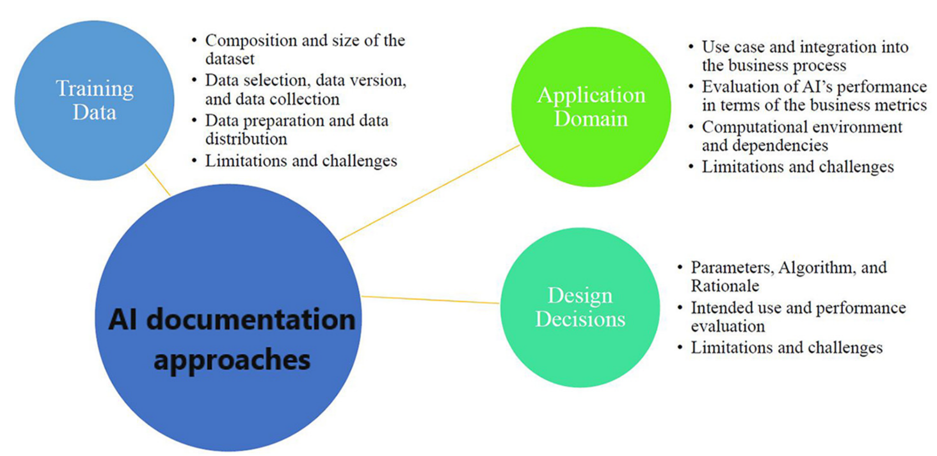
4.1.1 Composition and size of the data set. First, literature describes tools and methods for documenting data set size and composition. Guidelines help researchers with providing this information via summary statistics and visualizations (Gebru *et al.*, 2021; Gundersen *et al.*, 2018; Holland *et al.*, 2018; Isdahl and Gundersen, 2019; Mitchell *et al.*, 2019; Mora-Cantallops *et al.*, 2021; Rostamzadeh *et al.*, 2022; Schelter *et al.*, 2017). In addition,

Table 1 Dimensions for the review

Dimension	*Question that is answered*
Requirement that is addressed	Which one of the requirements is addressed?
Documentation method used	How does the paper document the AI? Through a guideline or an automated software tool?
Challenges and limitations of the documentation method	What challenges or limitations are discussed in the paper?

Source: Created by the author

Figure 1 Structure of the results



Source: Created by the author

documenting unbiasedness concerning protected attributes and testing for biases and fairness is enabled (Arnold *et al.*, 2019; Gebru *et al.*, 2021).

Software significantly simplifies creating and documenting training data set composition and size. Tools for computing and recording summary statistics and metadata are available (Alberti *et al.*, 2019; Gundersen *et al.*, 2018; Holland *et al.*, 2018; Isdahl and Gundersen, 2019; Schelter *et al.*, 2017; Wibisono *et al.*, 2014), along with visualization tools (Beg *et al.*, 2021; Souza *et al.*, 2019).

However, documenting unstructured data like audio, images or text is more complex. Few guidelines exist for data such as images (Miceli *et al.*, 2021), text (Bender and Friedman, 2018) and speech/audio (Papakyriakopoulos *et al.*, 2023; Srinivasan *et al.*, 2021). While some guidelines detail data set content (Papakyriakopoulos *et al.*, 2023; Srinivasan *et al.*, 2021), documentation often relies on metadata, focusing on demographics and data set variety (Bender and Friedman, 2018; Miceli *et al.*, 2021; Papakyriakopoulos *et al.*, 2023; Srinivasan *et al.*, 2021).

4.1.2 Data selection, version and collection. Second, the reviewed publications enable the documentation of information on the data selection, data version and data collection of the training data. Guidelines and checklists from different fields empower researchers to track data collection and selection rationale (Artrith *et al.*, 2021; Bender and Friedman, 2018; Hutchinson *et al.*, 2021; Isdahl and Gundersen, 2019; Rostamzadeh *et al.*, 2022; Rule *et al.*, 2019; Vasey *et al.*, 2022; Walsh *et al.*, 2021). Documentation guidelines also offer methods for recording how and when data was collected (Artrith *et al.*, 2021; Gebru *et al.*, 2021; Hutchinson *et al.*, 2021; Norgeot *et al.*, 2020; Srinivasan *et al.*, 2021). In the collection of unstructured data such as images or audio data, special attention needs to be paid to the documentation of the exact tools and mechanisms used for the collection of the data (Papakyriakopoulos *et al.*, 2023; Srinivasan *et al.*, 2021), since different tools may introduce different errors or biases into the data set. Legal aspects such as questions relating to the legality of collecting personal data and ownership rights to images and music can also be documented using existing guidelines (Artrith *et al.*, 2021; Gebru *et al.*, 2021; Hutchinson *et al.*, 2021; Norgeot *et al.*, 2020; Srinivasan *et al.*, 2021). Ethical considerations during data collection can also be documented (Gebru *et al.*, 2021; Mohammad, 2021). Guidelines have also been proposed for documenting crowd-sourced or crowd-annotated data (Diaz *et al.*, 2022). Various guides provide methods for recording the motivation and intended use of data sets (Gebru *et al.*, 2021; Mitchell *et al.*, 2019; Papakyriakopoulos *et al.*, 2023; Srinivasan *et al.*, 2021), and if the data set is representative and suitable for specific use cases (Cobbe *et al.*, 2021; Holland *et al.*, 2018; Hutchinson *et al.*, 2021; Norgeot *et al.*, 2020; Papakyriakopoulos *et al.*, 2023; Rostamzadeh *et al.*, 2022).

As data sets evolve, it is vital to document the data version used in AI training as well as whether and by whom the data set was maintained since a previous version of the data set was released (Artrith *et al.*, 2021; Holland *et al.*, 2018; Papakyriakopoulos *et al.*, 2023; Srinivasan *et al.*, 2021; Stodden and Miguez, 2013). Also, guidelines exist for making the decommission of data sets transparent (Luccioni *et al.*, 2022). Software tools can help with automating the documentation of data sources and storage locations (Holland *et al.*, 2018; Souza *et al.*, 2019), streamlining the documentation process and conserving researchers' time and effort.

4.1.3 Data preparation and distribution. Thirdly, there are established methods for documenting data cleaning, labeling procedures, feature generation and other methods for preparing and maintaining data sets. These methods include checklists, guidelines and specific software (Artrith *et al.*, 2021; Cobbe *et al.*, 2021; Gebru *et al.*, 2021; Mitchell *et al.*, 2019; Norgeot *et al.*, 2020; Vartak *et al.*, 2016; Vasey *et al.*, 2022; Walsh *et al.*, 2021). A lot of emphasis is given to the documentation of the data labeling (Diaz *et al.*, 2022; Gebru *et al.*, 2021; Papakyriakopoulos *et al.*, 2023), the calculation and selection of features (Arnold *et al.*, 2019; Cobbe *et al.*, 2021; Mitchell *et al.*, 2019) and the correction of unwanted biases or errors in the data set (Arnold *et al.*, 2019; Papakyriakopoulos *et al.*, 2023). Particular attention needs to be paid to the documentation of the data preparation of unstructured data such as audio or images (Miceli *et al.*, 2021; Papakyriakopoulos *et al.*, 2023; Srinivasan *et al.*, 2021) because unstructured data sets may contain data-type specific errors such as background noise or blurs. In addition, attention is paid to the documentation of any other data cleaning steps (Arnold *et al.*, 2019; Artrith *et al.*, 2021; Cobbe *et al.*, 2021; Mitchell *et al.*, 2019).

Another relevant topic is the question of whether and how the training data will be distributed and shared with other researchers and companies. Some guidelines only ask developers to document information on whether and to whom data will be distributed need to be answered (Gebru *et al.*, 2021; Papakyriakopoulos *et al.*, 2023; Srinivasan *et al.*, 2021) and emphasize that some copyrighted material may only be held privately or published under a restrictive license (Papakyriakopoulos *et al.*, 2023; Srinivasan *et al.*, 2021), whereas other publications actively advocate for and enable the sharing of training data under an open license (Gundersen *et al.*, 2018; Heil *et al.*, 2021; Rule *et al.*, 2019; Stodden and Miguez, 2013; Walsh *et al.*, 2021). Another interesting aspect in the context of computer generated data that can and should be documented is the question of whether generated data (i.e. music, text and images) will be distributed under the protection of strict intellectual property restrictions or under lesser protection (Srinivasan *et al.*, 2021).

4.1.4 Limitations. Despite numerous papers on data documentation methods, several challenges persist. Researchers cite data distribution restrictions like privacy and intellectual property laws as barriers (Heil *et al.*, 2021; Norgeot *et al.*, 2020; Stodden and Miguez, 2013). Solutions include using synthetic training data (Holland *et al.*, 2018; Srinivasan *et al.*, 2021) or obtaining data distribution consent (Gebru *et al.*, 2021). Noncommunication about private data sets' depreciation also poses a challenge (Luccioni *et al.*, 2022), impacting developers' adaptation needs. Extracting information from legacy systems for automatic documentation is also considered difficult (Schelter *et al.*, 2017). Ethical decisions cannot rely solely on documentation, human judgment is essential (Papakyriakopoulos *et al.*, 2023). Many guidelines, developed by researchers, offer limited practical support for AI developers (Srinivasan *et al.*, 2021). Not all guidelines are standalone; some are extensions, requiring complementary use (Papakyriakopoulos *et al.*, 2023; Srinivasan *et al.*, 2021). Finally, evolving definitions of inappropriate bias necessitate regular AI and documentation updates for different groups (Schramowski *et al.*, 2022).

4.2 Documentation of the application domain

At the same time, the literature also introduces methods for documenting the AI's application domain.

4.2.1 Introduction of the use case and integration of the artificial intelligence into the business process. The literature presents several methods to document the use case in which the AI application is used and how the AI application is integrated into the institution's processes. The questions that are being asked fall into one of four categories. First, multiple guidelines and checklists include specific questions on the use case in which the AI will be deployed and to justify the use of an AI in this context (Cobbe *et al.*, 2021; Liu *et al.*, 2020; Miceli *et al.*, 2021; Norgeot *et al.*, 2020; Rivera *et al.*, 2020). In medical studies, for instance, such a description would coincide with the description of the overall study design and justification for the study (Norgeot *et al.*, 2020; Rivera *et al.*, 2020). Second, the documentation of the objective of the AI application is to be documented in several guidelines. This is often done through a research question, business requirements or hypotheses (Cobbe *et al.*, 2021; Gundersen *et al.*, 2018; Isdahl and Gundersen, 2019; Liu *et al.*, 2020; Rivera *et al.*, 2020). Third, researchers support the documentation of the actions that are being taken based on predictions of the AI and whether these actions can be overruled or skipped by a human employee (Cobbe *et al.*, 2021; Liu *et al.*, 2020;

Rivera *et al.*, 2020; Vasey *et al.*, 2022). Fourth, several guidelines enable researchers and developers to document the interaction between users and the AI and specify the preferred characteristics of the intended user (Liu *et al.*, 2020; Mitchell *et al.*, 2019; Rivera *et al.*, 2020; Vasey *et al.*, 2022).

4.2.2 Evaluation of artificial intelligence's performance in terms of the business metrics. In addition to methods for documenting information on the use case, methods for documenting the performance of the AI in terms of relevant business metrics exist. The translation of the AI performance to relevant business metrics differs from the technical evaluation of the AI performance in the sense that it focusses primarily on the business impact of the predictions of the AI. Researchers are empowered to document instructions on how the predictions made by the AI should be translated into use case specific metrics and instructions (Norgeot *et al.*, 2020) and to evaluate how well the AI does in terms of the metrics of the use case (Vasey *et al.*, 2022). In addition, researchers and AI developers are given instructions on how to document methods for detecting and mitigating physical harm and ethical risks of the AI application (Cobbe *et al.*, 2021; Mohammad, 2021; Rivera *et al.*, 2020; Vasey *et al.*, 2022), whereas other guidelines only show researchers and developers how to document potential damage of the AI (Liu *et al.*, 2020; Mitchell *et al.*, 2019).

4.2.3 Computational environment and dependencies. Several documentation methods have been proposed to ensure that the computational environment can be documented. First, guidelines and checklists can be used to document the software tools, libraries and the hardware resources that went into the training, operation and maintenance of the AI application (Cobbe *et al.*, 2021; Dodge *et al.*, 2019; Heil *et al.*, 2021; Rule *et al.*, 2019; Stodden and Miguez, 2013). For this purpose, tools for the automatic documentation of the software and hardware environment exist as well (Isdahl and Gundersen, 2019). Second, guidelines suggest that researchers document the version of the AI used in the specific use case (Rivera *et al.*, 2020; Rule *et al.*, 2019; Stodden and Miguez, 2013). Together, the documentation of the computational environment and the dependencies can simplify tracing errors made by the AI.

4.2.4 Limitations. Researchers identify several challenges in documenting AI application domains. First, current methods often overlook documenting AI applications' usability and actual usage, risking unintended user application (Arnold *et al.*, 2019; Isdahl and Gundersen, 2019). Second, technical challenges in AI application creation and maintenance make documentation complex. AI engineers may prioritize integration with IT systems over integration with business processes (Isdahl and Gundersen, 2019). Third, using external software components demands separate documentation for reproducibility and fairness (Arnold *et al.*, 2019; Heil *et al.*, 2021), and computational requirements may impede reproducibility (Heil *et al.*, 2021).

4.3 Documenting the design decisions made during artificial intelligence development

Finally, literature encompassing guidelines, checklists and software for documenting design decisions in AI development exists.

4.3.1 Parameters, algorithm and rationale behind design decisions. First, methods for documenting the algorithm, parameters and optimization procedures have been found. Several guidelines ask researchers and data scientists to document the model type, algorithms, parameters, features and the parameter tuning process (Dodge *et al.*, 2019; Isdahl and Gundersen, 2019; Mitchell *et al.*, 2019; Rule *et al.*, 2019; Vasey *et al.*, 2022; Walsh *et al.*, 2021). Additionally, methods for preventing biases in the decisions of the AI can be documented (Arnold *et al.*, 2019; Cobbe *et al.*, 2021; Mitchell *et al.*, 2019; Walsh *et al.*, 2021). In addition, papers acknowledge that AI applications are maintained and changed and state that the version can be documented using the version number of the AI

(Alberti *et al.*, 2018; Alberti *et al.*, 2019; Mitchell *et al.*, 2019) or by making the code used for training the AI publicly accessible (Heil *et al.*, 2021; Stodden and Miguez, 2013; Walsh *et al.*, 2021; Wibisono *et al.*, 2014). Additionally, software tools for automatically documenting the chosen parameters and decisions of the researchers and a discussion of possible alternate workflows have been proposed (Alberti *et al.*, 2019; Beg *et al.*, 2021; Mora-Cantallops *et al.*, 2021; Schelter *et al.*, 2017; Wang *et al.*, 2021).

4.3.2 Intended use and technical performance evaluation. Second, the indented use case and the technical performance of the AI model can be documented. Documentation can include the AI model's purpose, intended and out-of-scope use cases and use-case-specific checklists and general guidelines (Arnold *et al.*, 2019; Cobbe *et al.*, 2021; Crisan *et al.*, 2022; Mitchell *et al.*, 2019). Many researchers support the documentation and justification of optimization, test metrics and technical performance results (Crisan *et al.*, 2022; Liu *et al.*, 2020; Norgeot *et al.*, 2020; Rule *et al.*, 2019; Vasey *et al.*, 2022; Walsh *et al.*, 2021). Guidelines often support the comparison of AI performance to state-of-the-art models and documenting model robustness concerning parameter changes (Artrith *et al.*, 2021; Norgeot *et al.*, 2020; Walsh *et al.*, 2021). Publications pave the way for detailing data set splitting methods, distribution and interdependencies among train, test and validation sets, and saving data set copies (Arnold *et al.*, 2019; Artrith *et al.*, 2021; Dodge *et al.*, 2019; Norgeot *et al.*, 2020; Walsh *et al.*, 2021). In addition, visualizations of metrics and data set histories can be created and documented using available tools (Souza *et al.*, 2019; Vartak *et al.*, 2016).

4.3.3 Limitations. While numerous design decisions in AI application development can be documented, a hurdle persists. Researchers are concerned about exposing business-sensitive information and potentially violating intellectual property law (Norgeot *et al.*, 2020; Stodden and Miguez, 2013). They contend that these factors may restrict scientists and developers from disclosing certain design aspects in the documentation.

5. Discussion

While the paper offers substantial theoretical and practical insights, a comparison of AI documentation approaches with requirements exposes key challenges.

5.1 Creation of documentation guidelines that satisfy all documentation requirements

First, the analysis shows a gap between existing AI documentation practices and regulatory and industry requirements. Current methods often focus on specific requirements areas, like training data (Gebru *et al.*, 2021) or model development (Mitchell *et al.*, 2019), or are limited to certain industries or research areas (Liu *et al.*, 2020; Rivera *et al.*, 2020). Researchers suggest that multiple guidelines need merging to adequately document even a single requirement (Papakyriakopoulos *et al.*, 2023; Srinivasan *et al.*, 2021). There is ambiguity on how to effectively combine these guidelines for auditor and regulator satisfaction. This fragmented approach may not fully meet the complex documentation needs, particularly in regulated environments and for AI applications with a high risk to society.

Another key challenge in AI documentation is the lack of focus on documenting governance and risk mitigation processes for user safety and security. Only a few studies emphasize documenting error detection, harm mitigation (Mitchell *et al.*, 2019; Mohammad, 2021; Rivera *et al.*, 2020; Vasey *et al.*, 2022) or quality assurance (Cobbe *et al.*, 2021). This is a significant challenge, since Bernstein *et al.* (2023) highlight AI's potential to mislead medical professionals, exemplifying AI's ability to trick trained professionals and to cause harm. Also, proper documentation of safety and risk management systems is vital for meeting regulatory and practitioner requirements (European Parliament, 2024; Königstorfer and

Thalmann, 2021; Krumay *et al.*, 2020). Research should assess if current guidelines for documenting safety, security and governance processes are sufficient to meet these regulatory and practitioner requirements.

Additionally, existing documentation approaches primarily cater to researchers, with none evaluated by governance experts or auditors, leaving companies and AI developers with inadequate guidance. This is crucial, as AI pose different challenges in practice than in laboratories (Hutchinson *et al.*, 2021; Miceli *et al.*, 2020). As a result, some risk factors associated with the AI may remain untransparent and could get overlooked, posing a significant risk to users and society as a whole.

5.2 Dealing with the evolving nature of artificial intelligence

Second, addressing new AI developments, like fine-tuning public pretrained models (Qinghua Lu *et al.*, 2023) for personalized applications, presents challenges. Pretrained models often lack thorough documentation, leading to development errors remaining unnoticed. Even peer-reviewed AI model papers often have critical flaws due to issues like improper feature selection or inadequate data cleaning (Kapoor and Narayanan, 2022). For instance, Northcutt *et al.* (2021) identified issues like inaccurate labels and duplicates in benchmark data sets, impacting results. Despite efforts to rectify label inaccuracies in ImageNet, ResNet-18 outperformed ResNet-50, contrary to published findings (Northcutt *et al.*, 2021). For many pretrained models, it is unclear whether these errors have been corrected. As a result, doubts persist about whether adequately documenting AI applications using pretrained models is feasible. This underscores the need for more research on documenting such AI applications and ensuring their governance, not just in research but also for other publicly available pretrained models.

In addition, the documentation of AI models that are frequently retrained has not been adequately addressed. Such models are regularly updated with new data (Zhu and Klabjan, 2021), making a static documentation inadequate due to constant changes in training data and model performance. Yet, current guidelines do not cover how to document these models and their retraining procedures. This gap highlights the need for more research on documenting frequently retrained AI models and their specific retraining processes.

Addressing the documentation of generative AI models, like large language models (LLMs), is also a notable challenge. Generative AI's ability to create content such as images and text makes it significantly different from traditional supervised or unsupervised AI models (AlDahoul *et al.*, 2023; Kirelli, 2023; Russell and Norvig, 2010). The distinct nature of tasks addressed by generative AI may require a different approach to documentation, underscoring the need for further research in documenting these advancements in AI.

5.3 Automatic documentation of artificial intelligence

Third, the use of automated tools for AI documentation offers new opportunities and challenges. Researchers and AI developers can now use software to automatically document training data and design decisions (Alberti *et al.*, 2019; Beg *et al.*, 2021; Mora-Cantallops *et al.*, 2021; Schelter *et al.*, 2017; Wang *et al.*, 2021). In addition, LLMs have shown success in describing and contextualizing code with minimal input (Sarsa *et al.*, 2022). This is promising for several reasons. First, automation can save significant time, leading to monetary savings (Ashurst *et al.*, 2022; Vasey *et al.*, 2022; Wang *et al.*, 2021). Wang *et al.* (2021) discovered that their solution could fully automate 45% of documentation tasks while suggesting that for an additional 41% of tasks, only minor adjustments needed to be made, significantly cutting down on employee time required for AI application documentation. This could save companies a significant amount of money. Second, with companies deploying multiple, frequently retraining AI models (Kashyap *et al.*, 2021), automation can make frequent documentation of these changes possible. This allows companies to enhance AI quality continuously, while meeting regulatory and social standards, thereby potentially gaining a competitive edge. Third, automated documentation can assist AI engineers in integrating AI applications into business processes, often neglected until late in the development process (Isdahl and Gundersen, 2019). Fourth, automated tools can improve the transparency and reproducibility of AI research. Researchers could make their work more transparent, and developers might document previously undocumented models, aiding error detection.

However, an evaluation of the effectiveness of automated documentation tools in practice is lacking, indicating a need for further investigation into their potential and limitations. While some tools report time savings in a lab setting (Wang *et al.*, 2021), it is uncertain if these translate into real cost savings or adaptability to changing requirements. In addition, deploying high-quality LLMs can be costly (Aryan *et al.*, 2023), potentially offsetting the benefits of saving time. Furthermore, LLMs often produce inaccurate or fabricated responses (Shi *et al.*, 2023; Zhang *et al.*, 2023), leading to potentially incorrect documentation. Convincing, yet erroneous AI predictions can mislead professionals, even against their correct intuition (Bernstein *et al.*, 2023), meaning that a detailed review of the created documentation may still be necessary. Researchers need to investigate how to ensure that the resulting documentation is correct, and create guidelines for when and how LLMs can and should be used for the documentation of AI applications. Also, this paper emphasizes the need to examine the time and cost benefits of automatic documentation tools.

6. Conclusion

This paper has explored the landscape of AI documentation, highlighting its critical role in the governance, efficacy and ethical deployment of AI applications. A literature review identified diverse methods and tools for documenting AI aspects, from training data to application domains. These findings highlight AI documentation's complexity and the challenges in developing comprehensive and transparent practices.

First, the paper identifies a notable gap between existing AI documentation practices and regulatory and industry requirements. Current methods are often specific to certain AI development aspects or industries, leading to a fragmented approach insufficient for complex, regulated environments. It highlights the need for comprehensive documentation guidelines covering all AI aspects, including safety, security, governance and societal impact.

Second, the findings emphasize the evolving AI landscape, including pretrained models, frequent retraining and generative models like LLMs. The rapid progress in AI introduces unique documentation challenges, with these technologies transcending traditional limits and adding new risks and governance complexities. The paper advocates for more research on effectively documenting these advanced AI models, given their significant impact on decision-making in critical areas.

Third, the paper explores automated AI documentation, like LLMs, revealing opportunities for efficient, accurate documentation. While automation may reduce time and effort, it raises concerns about documentation accuracy and reliability. The paper stresses the need to validate automated documentation, balancing efficiency with the necessity for accuracy and compliance.

In conclusion, this paper greatly enhances understanding of AI documentation, providing an in-depth analysis of current practices and pinpointing future research directions. As AI evolves and integrates into various sectors, robust, transparent and comprehensive documentation is increasingly essential. This research lays the groundwork for effective documentation standards that align with regulatory needs and promote trust and accountability in AI systems, guiding their responsible and ethical utilization.

The limitations of this paper include a limited scope of databases and journals consulted, as well as a language and geographic bias that may exclude relevant studies published in languages other than English or from diverse geographical regions. Also, the reliance on existing literature without considering tools and methods from practice for AI documentation is a limitation. In addition, the reliance on IS and medical literature can be seen as a limitation. Furthermore, regulatory requirements may evolve, requiring a reevaluation of the literature and additional research at a later date.

References

Alberti, M., Pondenkandath, V., Vögtlin, L., Würsch, M., Ingold, R. and Liwicki, M. (2019), "Improving reproducible deep learning workflows with deepdiva", *2019 6th Swiss Conference on Data Science (SDS)*, pp. 13-18.

Alberti, M., Pondenkandath, V., Würsch, M., Ingold, R. and Liwicki, M. (2018), "DeepDIVA: a highly-functional python framework for reproducible experiments", *2018 16th International Conference on Frontiers in Handwriting Recognition (ICFHR)*, pp. 423-428.

AlDahoul, N., Hong, J., Varvello, M. and Zaki, Y. (2023), "Exploring the potential of generative AI for the world wide web", arXiv preprint arXiv:2310.17370.

Alfaro, E., Bressan, M., Girardin, F., Murillo, J., Someh, I. and Wixom, B.H. (2019), "BBVA's data monetization journey", *MIS Quarterly Executive*, Vol. 18 No. 2, p. 117.

Appelbaum, D., Kogan, A. and Vasarhelyi, M.A. (2017), "Big data and analytics in the modern audit engagement: research needs", *Auditing: A Journal of Practice & Theory*, Vol. 36 No. 4, pp. 1-27.

Arnold, M., Bellamy, R.K.E., Hind, M., Houde, S., Mehta, S., Mojsilović, A., Nair, R., Ramamurthy, K. N., Olteanu, A. and Piorkowski, D. (2019), "FactSheets: increasing trust in AI services through supplier's declarations of conformity", *IBM Journal of Research and Development*, Vol. 63 Nos 4/5, p. 13.

Artrith, N., Butler, K.T., Coudert, F.-X., Han, S., Isayev, O., Jain, A. and Walsh, A. (2021), "Best practices in machine learning for chemistry", *Nature Chemistry*, Vol. 13 No. 6, pp. 505-508.

Aryan, A., Nain, A.K., McMahon, A., Meyer, L.A. and Sahota, H.S. (2023), "The costly dilemma: generalization, evaluation and cost-optimal deployment of large language models", arXiv preprint arXiv:2308.08061.

Ashurst, C., Hine, E., Sedille, P. and Carlier, A. (2022), "AI ethics statements: analysis and lessons learnt from NeurIPS broader impact statements", *2022 ACM Conference on Fairness, Accountability, and Transparency*, pp. 2047-2056.

Beg, M., Taka, J., Kluyver, T., Konovalov, A., Ragan-Kelley, M., Thiéry, N.M. and Fangohr, H. (2021), "Using Jupyter for reproducible scientific workflows", *Computing in Science & Engineering*, Vol. 23 No. 2, pp. 36-46.

Bender, E.M. and Friedman, B. (2018), "Data statements for natural language processing: toward mitigating system bias and enabling better science", *Transactions of the Association for Computational Linguistics*, Vol. 6, pp. 587-604.

Bernstein, M.H., Atalay, M.K., Dibble, E.H., Maxwell, A.W.P., Karam, A.R., Agarwal, S., Ward, R.C., Healey, T.T. and Baird, G.L. (2023), "Can incorrect artificial intelligence (AI) results impact radiologists, and if so, what can we do about it? A multi-reader pilot study of lung cancer detection with chest radiography", *European Radiology*, Vol. 33 No. 11, pp. 1-7.

Clements, P., Bachmann, F., Bass, L., Garlan, D., Ivers, J., Little, R., Merson, P., Nord, R. and Stafford, J. (2011), *Documenting Software Architectures: Views and Beyond*, Addison-Wesley Professional, Boston, MA.

Cobbe, J., Lee, M.S.A. and Singh, J. (2021), "Reviewable automated decision-making: a framework for accountable algorithmic systems", *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pp. 598-609.

Collins, G.S., Reitsma, J.B., Altman, D.G. and Moons, K.G.M. (2015), "Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement", *Annals of Internal Medicine*, Vol. 162 No. 1, pp. 55-63.

Crisan, A., Drouhard, M., Vig, J. and Rajani, N. (2022), "Interactive model cards: a human-centered approach to model documentation", *2022 ACM Conference on Fairness, Accountability, and Transparency*, pp. 427-439.

Diaz, M., Kivlichan, I., Rosen, R., Baker, D., Amironesei, R., Prabhakaran, V. and Denton, E. (2022), "Crowdworksheets: accounting for individual and collective identities underlying crowdsourced dataset annotation", *2022 ACM Conference on Fairness, Accountability, and Transparency*, pp. 2342-2351.

Dodge, J., Gururangan, S., Card, D., Schwartz, R. and Smith, N.A. (2019), "Show your work: improved reporting of experimental results", arXiv preprint arXiv:1909.03004.

Dressel, J. and Farid, H. (2018), "The accuracy, fairness, and limits of predicting recidivism", *Science Advances*, Vol. 4 No. 1, p. eaao5580.

Ellul, J., Pace, G., McCarthy, S., Sammut, T., Brockdorff, J. and Scerri, M. (2021), "Regulating artificial intelligence: a technology regulator's perspective", Proceedings of the Eighteenth International Conference on Artificial Intelligence and Law, pp. 190-194.

European Parliament (2024), "Proposal for a regulation of the European parliament and of the council laying down harmonised rules on artificial intelligence (artificial intelligence act) and amending certain union legislative acts", available at: https://data.consilium.europa.eu/doc/document/ST-5662-2024-INIT/en/pdf (accessed 19 March 2024).

Garousi, G., Garousi-Yusifoğlu, V., Ruhe, G., Zhi, J., Moussavi, M. and Smith, B. (2015), "Usage and usefulness of technical software documentation: an industrial case study", *Information and Software Technology*, Vol. 57, pp. 664-682.

Gashi, M., Vuković, M., Jekic, N., Thalmann, S., Holzinger, A., Jean-Quartier, C. and Jeanquartier, F. (2022), "State-of-the-Art explainability methods with focus on visual analytics showcased by glioma classification", *BioMedInformatics*, Vol. 2 No. 1, pp. 139-158.

Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J.W., Wallach, H., Daumé, H., III. and Crawford, K. (2021), "Datasheets for datasets", *Communications of the ACM*, Vol. 64 No. 12, pp. 86-92.

Gundersen, O.E. and Kjensmo, S. (2018), "State of the art: reproducibility in artificial intelligence", *Proceedings of the AAAI Conference on Artificial Intelligence*.

Gundersen, O.E., Gil, Y. and Aha, D.W. (2018), "On reproducible AI: towards reproducible research, open science, and digital scholarship in AI publications", *AI Magazine*, Vol. 39 No. 3, pp. 56-68.

Heil, B.J., Hoffman, M.M., Markowetz, F., Lee, S.-I., Greene, C.S. and Hicks, S.C. (2021), "Reproducibility standards for machine learning in the life sciences", *Nature Methods*, Vol. 18 No. 10, pp. 1-4.

Heinrichs, B. (2022), "Discrimination in the age of artificial intelligence", *AI & Society*, Vol. 37 No. 1, pp. 143-154.

Holland, S., Hosny, A., Newman, S., Joseph, J. and Chmielinski, K. (2018), "The dataset nutrition label: a framework to drive higher data quality standards", arXiv preprint arXiv:1805.03677.

Hutchinson, B., Smart, A., Hanna, A., Denton, E., Greer, C., Kjartansson, O., Barnes, P. and Mitchell, M. (2021), "'Towards accountability for machine learning datasets: practices from software engineering and infrastructure", *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pp. 560-575.

Isdahl, R. and Gundersen, O.E. (2019), "Out-of-the–box reproducibility: a survey of machine learning platforms", *2019 15th international conference on eScience (eScience)*, pp. 86-95.

ISO (2019), "ISO/IEC/IEEE 15289:2019(en) - Systems and software engineering–content of life-cycle information items (documentation): ISO [online]", available at: www.iso.org/standard/74909.html (accessed 17 December 2023).

Kapoor, S. and Narayanan, A. (2022), "Leakage and the reproducibility crisis in ML-based Science", arXiv preprint arXiv:2207.07048.

Kashyap, S., Morse, K.E., Patel, B. and Shah, N.H. (2021), "A survey of extant organizational and computational setups for deploying predictive models in health systems", *Journal of the American Medical Informatics Association*, Vol. 28 No. 11, pp. 2445-2450.

Kirelli, Y. (2023), “Analysis of factors affecting common use of generative artificial intelligence-based tools by machine learning methods”, *International Journal of Computational and Experimental Science and Engineering*, Vol. 9 No. 3, pp. 233-237.

Königstorfer, F. and Thalmann, S. (2021), “Software documentation is not enough! Requirements for the documentation of AI”, *Digital Policy, Regulation and Governance*, Vol. 23 No. 5, pp. 475-488.

Königstorfer, F. and Thalmann, S. (2022), “AI documentation: a path to accountability”, *Journal of Responsible Technology*, Vol. 11, pp. 100043-100053.

Königstorfer, F., Haberl, A., Kowald, D., Ross-Hellauer, T. and Thalmann, S. (2024), “Black box or open science? A study on reproducibility in AI development papers”, *57th Annual Hawaii International Conference on System Sciences, HICSS 2024*.

Krumay, B., Koch, S. and Winkler, M. (2020), “Einhaltung von informationssicherheitsvorschriften durch MitarbeiterInnen: faktoren und maßnahmen”, *Wirtschaftsinformatik (Zentrale Tracks)*, pp. 1294-1308.

Kupfer, C., Prassl, R., Fleiß, J., Malin, C., Thalmann, S. and Kubicek, B. (2023), “Check the box! How to deal with automation bias in AI-based personnel selection”, *Frontiers in Psychology*, Vol. 14, p. 1118723.

Liu, X., Rivera, S.C., Moher, D., Calvert, M.J. and Denniston, A.K. (2020), “Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension”, *BMJ*, Vol. 370, pp. e537-e548.

Lu, Q., Zhu, L., Xu, X., Xing, Z. and Whittle, J. (2023), “A taxonomy of foundation model based systems for Responsible-AI-by-Design”.

Luccioni, A.S., Corry, F., Sridharan, H., Ananny, M., Schultz, J. and Crawford, K. (2022), “A framework for deprecating datasets: standardizing documentation identification and communication”, *2022 ACM Conference on Fairness, Accountability, and Transparency*, pp. 199-212.

Makridakis, S. (2017), “The forthcoming artificial intelligence (AI) revolution: its impact on society and firms”, *Futures*, Vol. 90, pp. 46-60.

Miceli, M., Schuessler, M. and Yang, T. (2020), “Between subjectivity and imposition: power dynamics in data annotation for computer vision”, *Proceedings of the ACM on Human-Computer Interaction*, Vol. 4 No. CSCW2, pp. 1-25.

Miceli, M., Yang, T., Naudts, L., Schuessler, M., Serbanescu, D. and Hanna, A. (2021), “Documenting computer vision datasets: an invitation to reflexive data practices”, *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pp. 161-172.

Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I.D. and Gebru, T. (2019), “Model cards for model reporting”, *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pp. 220-229.

Mohammad, S.M. (2021), “Ethics sheets for AI tasks”, arXiv preprint arXiv:2107.01183.

Mora-Cantallops, M., Sánchez-Alonso, S., García-Barriocanal, E. and Sicilia, M.-A. (2021), “Traceability for trustworthy AI: a review of models and tools”, *Big Data and Cognitive Computing*, Vol. 5 No. 2, p. 20.

Ndlovu, S.L. and Kyobe, M.E. (Eds) (2016), *Challenges of CoBIT 5 IT Governance Framework Migration*, In Proceedings of the International Conference on Information Resources Management (CONFIRM).

Norgeot, B., Quer, G., Beaulieu-Jones, B.K., Torkamani, A., Dias, R., Gianfrancesco, M., Arnaout, R., Kohane, I.S., Saria, S. and Topol, E. (2020), “Minimum information about clinical artificial intelligence modeling: the MI-CLAIM checklist”, *Nature Medicine*, Vol. 26 No. 9, pp. 1320-1324.

Northcutt, C.G., Athalye, A. and Mueller, J. (2021), “Pervasive label errors in test sets destabilize machine learning benchmarks”, arXiv preprint arXiv:2103.14749.

Papakyriakopoulos, O., Choi, A.S.G., Thong, W., Zhao, D., Andrews, J., Bourke, R., Xiang, A. and Koenecke, A. (2023), “Augmented datasheets for speech datasets and ethical decision-making”, Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency, pp. 881-904.

Patton, M.Q. (2014), *Qualitative Research & Evaluation Methods: Integrating Theory and Practice*, Sage publications, Thousand Oaks, California, USA.

Rivera, S.C., Liu, X., Chan, A.-W., Denniston, A.K. and Calvert, M.J. (2020), “Guidelines for clinical trial protocols for interventions involving artificial intelligence: the SPIRIT-AI extension”, *BMJ*, Vol. 370, pp. e549-e560.

Rostamzadeh, N., Mincu, D., Roy, S., Smart, A., Wilcox, L., Pushkarna, M., Schrouff, J., Amironesei, R., Moorosi, N. and Heller, K. (2022), "Healthsheet: development of a transparency artifact for health datasets", arXiv preprint arXiv:2202.13028.

Rowley, J. and Slack, F. (2004), "Conducting a literature review", *Management Research News*, Vol. 27 No. 6, pp. 31-39.

Rule, A., Birmingham, A., Zuniga, C., Altintas, I., Huang, S.-C., Knight, R., Moshiri, N., Nguyen, M.H., Rosenthal, S.B. and Pérez, F. (2019), "Ten simple rules for writing and sharing computational analyses in Jupyter Notebooks", Public Library of Science.

Russell, S. and Norvig, P. (2010), *Artificial Intelligence: A Modern Approach*, 3rd ed., Pearson Education, Harlow.

Sadek, M., Kallina, E., Bohné, T., Mougenot, C., Calvo, R.A. and Cave, S. (2024), "Challenges of responsible AI in practice: scoping review and recommended actions", *AI & Society*, pp. 1-17.

Sarsa, S., Denny, P., Hellas, A. and Leinonen, J. (2022), "Automatic generation of programming exercises and code explanations using large language models", *Proceedings of the 2022 ACM Conference on International Computing Education Research-Volume 1*, pp. 27-43.

Schelter, S., Boese, J.-H., Kirschnick, J., Klein, T. and Seufert, S. (2017), "Automatically tracking metadata and provenance of machine learning experiments", Machine Learning Systems Workshop at NIPS, pp. 27-29.

Schramowski, P., Tauchmann, C. and Kersting, K. (2022), "Can machines help Us answering question 16 in datasheets, and in turn reflecting on inappropriate content?", *2022 ACM Conference on Fairness, Accountability, and Transparency*.

Shi, F., Chen, X., Misra, K., Scales, N., Dohan, D., Chi, E.H., Schärli, N. and Zhou, D. (2023), "Large language models can be easily distracted by irrelevant context", International Conference on Machine Learning, pp. 31210-31227.

Simonsson, M., Johnson, P. and Ekstedt, M. (2010), "The effect of IT governance maturity on IT governance performance", *Information Systems Management*, Vol. 27 No. 1, pp. 10-24.

Souza, R., Azevedo, L., Lourenço, V., Soares, E., Thiago, R., Brandão, R., Civitarese, D., Brazil, E., Moreno, M., Valduriez, P. and Mattoso, M. (2019), "Provenance data in the machine learning lifecycle in computational science and engineering", *2019 IEEE/ACM Workflows in Support of Large-Scale Science (WORKS)*, pp. 1-10.

Srinivasan, R., Denton, E., Famularo, J., Rostamzadeh, N., Diaz, F. and Coleman, B. (2021), "Artsheets for art datasets", Thirty-fifth conference on neural information processing systems datasets and benchmarks track.

Stodden, V. and Miguez, S. (2013), "Best practices for computational science: software infrastructure and environments for reproducible and extensible research", Available at SSRN 2322276.

Sun, C., Shrivastava, A., Singh, S. and Gupta, A. (Eds) (2017), *Revisiting Unreasonable Effectiveness of Data in Deep Learning Era*, In Proceedings of the IEEE International Conference on Computer Vision (ICCV), IEEE, pp. 843-852, doi: 10.1109/ICCV.2017.97.

Vartak, M., Subramanyam, H., Lee, W.-E., Viswanathan, S., Husnoo, S., Madden, S. and Zaharia, M. (2016), "'ModelDB: a system for machine learning model management", *Proceedings of the Workshop on Human-In-the-Loop Data Analytics*, pp. 1-3.

Vasey, B., Nagendran, M., Campbell, B., Clifton, D.A., Collins, G.S., Denaxas, S., Denniston, A.K., Faes, L., Geerts, B., Ibrahim, M. and Liu, X. (2022), "Reporting guideline for the early-stage clinical evaluation of decision support systems driven by artificial intelligence: DECIDE-AI", *Nature Medicine*, Vol. 28 No. 5, pp. 924-933.

Wachter, S. and Mittelstadt, B. (2019), "A right to reasonable inferences: re-thinking data protection law in the age of big data and AI", *Colum. Bus. L. Rev*, p. 494.

Walsh, I., Fishman, D., Garcia-Gasulla, D., Titma, T., Pollastri, G., Harrow, J., Psomopoulos, F.E. and Tosatto, S.C.E. (2021), "DOME: recommendations for supervised machine learning validation in biology", *Nature Methods*, Vol. 18 No. 11, pp. 1-6.

Wang, A.Y., Wang, D., Drozdal, J., Muller, M., Park, S., Weisz, J.D., Liu, X., Wu, L. and Dugan, C. (2021), "Themisto: towards automated documentation generation in computational notebooks", arXiv preprint arXiv:2102.12592.

Webster, J. and Watson, R.T. (2002), “Analyzing the past to prepare for the future: writing a literature review”, *Mis Quarterly*, pp. 13-23.

Wibisono, A., Bloem, P., de Vries, G.K.D., Groth, P., Belloum, A. and Bubak, M. (Eds) (2014), *Generating Scientific Documentation for Computational Experiments Using Provenance*, in Ludäscher, B. and Plale, B. (Eds), Provenance and Annotation of Data and Processes: 5th International Provenance and Annotation Workshop, IPAW 2014, Cologne, Germany, 9-13 June, 2014: Revised Selected Papers, Springer, pp. 168-179, doi: 10.1007/978-3-319-16462-5_13.

Winter, J.S. and Davidson, E. (2019), “Governance of artificial intelligence and personal health information”, *Digital Policy, Regulation and Governance*, Vol. 21 No. 3, pp. 280-290.

Zhang, M., Press, O., Merrill, W., Liu, A. and Smith, N.A. (2023), “How language model hallucinations can snowball”, arXiv preprint arXiv:2305.13534.

Zhu, X. and Klabjan, D. (2021), “Continual neural network model retraining”, *2021 IEEE International Conference on Big Data (Big Data)*, pp. 1163-1171.

Corresponding author

Florian Königstorfer can be contacted at: florian.koenigstorfer@edu.uni-graz.at