The CLeAR Documentation Framework for AI Transparency

Recommendations for Practitioners and Context for Policymakers

MAY 2024

The views expressed in Shorenstein Center Discussion Papers are those of the author(s) and do not necessarily reflect those of Harvard Kennedy School, Harvard University, or the organizations and institutions with whom the authors are affiliated.

Discussion Papers have not undergone formal review and approval. Such papers are included in this series to elicit feedback and to encourage debate on important issues and challenges in media, politics and public policy. Copyright belongs to the author(s). Papers may be downloaded for personal use only.

Authors

Kasia Chmielinski* (Data Nutrition Project & Harvard University)
Sarah Newman* (Data Nutrition Project & Harvard University)
Chris N. Kranzinger* (Data Nutrition Project)
Michael Hind (IBM Research)
Jennifer Wortman Vaughan (Microsoft Research)
Margaret Mitchell (Hugging Face)
Julia Stoyanovich (New York University)
Angelina McMillan-Major (independent researcher)
Emily McReynolds (independent researcher)
Kathleen Esfahany (Data Nutrition Project & Harvard University)
Mary L. Gray (Microsoft Research, Indiana University & Harvard University)
Maui Hudson (TKRI, University of Waikato)
Audrey Chang (Data Nutrition Project & Harvard University)

*Joint first authors

HARVARD Kennedy School
SHORENSTEIN CENTER
on Media, Politics and Public Policy

Abstract

This report introduces the CLeAR (**C**omparable, **L**egible, **A**ctionable, and **R**obust) Documentation Framework to offer guiding principles for AI documentation. The framework is designed to help practitioners and others in the AI ecosystem consider the complexities and tradeoffs required when developing documentation for datasets, models, and AI systems (which contain one or more models, and often other software components). Documentation of these elements is crucial and serves several purposes, including: (1) Supporting responsible development and use, as well as mitigation of downstream harms, by providing transparency into the design, attributes, intended use, and shortcomings of datasets, models, and AI systems; (2) Motivating dataset, model, or AI system creators and curators to reflect on the choices they make; and (3) Facilitating dataset, model, and AI system evaluation and auditing. We assert that documentation should be mandatory in the creation, usage, and sale of datasets, models, and AI systems.

This framework was developed with the expertise and perspective of a team that has worked at the forefront of AI documentation across both industry and the research community. As the need for documentation in machine learning and AI becomes more apparent and the benefits more widely acknowledged, we hope it will serve as a guide for future AI documentation efforts as well as context and education for regulators' efforts toward AI documentation requirements.

Table of Contents

Introduction

Machine learning models are designed to identify patterns and make predictions based on statistical analysis of data. Modern artificial intelligence (AI) systems are often made up of one or more machine learning models combined with other components. While the statistical foundations of modern machine learning are decades old [1,2], technological developments in the last decade have led to both more widespread deployment of AI systems and an increasing awareness of the risks associated with them. The use of statistical analysis to make predictions is prone to replicate already existing biases and even introduce new ones [3], causing issues like racial bias in medical applications [4,5], gender bias in finance [6], and disability bias in public benefits [7]. AI systems can also lead to other (intentional as well as unintentional) harms [8,9], including the production of misinformation [10–13] and the sharing or leaking of private information [14].

This is a pivotal moment to think about the social impact of AI, including its risks and harms, and the implementation of accountability and governance more broadly.

With growing national and international attention to AI regulation [15,16], rights-based principles [17–21], data equity [22,23], and risk mitigation [24], this is a pivotal moment to think about the social impact of AI, including its risks and harms, and the implementation of accountability and governance more broadly. Most proposed AI regulation mandates some level of **transparency**, as transparency is crucial for addressing the ways in which AI systems impact people. This is because transparency is a foundational, extrinsic value—a means for other values to be realized. Applied to AI development, transparency can enhance accountability by making it clear who is responsible for which kinds of system behavior. This can lessen the amount of time it takes to stop harms from proliferating once they are identified, and provides legal recourse when people are harmed. Transparency guides developers towards **non-discrimination** in deployed systems, as it encourages testing for it and being transparent about evaluation results. Transparency enables **reproducibility**, as details provided can then be followed by others. This in turn incentivizes **integrity** and **scientific rigor** in claims made by AI developers and deployers and improves the **reliability** of systems. And transparency around how an AI system works can foster appropriate levels of trust from users and enhance **human agency.**

Transparency can be realized, in part, by providing information about how the data used to develop and evaluate the AI system was collected and processed, how AI models were built, trained, and fine-tuned, and how models and systems were evaluated and deployed [25,26]. Towards this end, documentation has emerged as an essential component of AI transparency [27] and a foundation for responsible AI development.

Although AI documentation may seem like a simple concept, determining what kinds of information should be documented and how this information should be conveyed is challenging in practice. Over the past decade, several concrete approaches, as well as evaluations of these approaches, have been proposed for documenting datasets [28–41], models [42–46], and AI systems [47–55]. Together, this body of work supplies a rich set of lessons learned.

The CLeAR Documentation Framework is designed to help practitioners and policymakers understand what principles should guide the process and content of AI documentation and how to create such documentation.

Based on these lessons, this paper introduces the **CLeAR Documentation Framework**, designed to help practitioners and policymakers understand what principles should guide the process and content of AI documentation and how to create such documentation. The CLeAR Documentation Framework introduces four principles for documentation and offers definitions, recommends approaches, explains tradeoffs, highlights open questions (Table 1), and helps guide the implementation of documentation (Table 2). It builds on and is aligned with previous principles-based frameworks for documentation [44,56].

At a high level, the **CLeAR Principles** state that documentation should be:

- **Comparable:** Able to be compared; having similar components to documentation of other datasets, models, or systems to permit or suggest comparison; enabling comparison by following a discrete, well-defined format in process, content, and presentation.
- **Legible:** Able to be read and understood; clear and accessible for the intended audience.
- **Actionable:** Able to be acted on; having practical value, useful for the intended audience.
- **Robust:** Able to be sustained over time; up to date.

A healthy documentation ecosystem requires the participation of both practitioners and policymakers. While both audiences benefit from understanding current approaches, principles, and tradeoffs, practitioners face more practical challenges of implementation, while policymakers need to establish structures for accountability that include and also extend *beyond* documentation, as documentation is necessary but not sufficient for accountability when it comes to complex technological systems. We offer guidance for both audiences.

Documentation, like most responsible AI practices, is an iterative process. We cannot expect a single approach to fit all use cases, and more work needs to be done to establish approaches that are effective in different contexts. We hope this document will provide a basis for this work.

Why now?

While experts have been alerting the public about the shortcomings of AI systems for some time [3,57,58], we have recently witnessed an unprecedented surge in public awareness and concern regarding AI. This recent attention can be attributed in part to the development of large language models by prominent tech companies, such as Google's LaMDA [59] and Gemini [60], Google DeepMind's Sparrow [61], Meta AI's Galactica [62] and LLaMa 2 [63], and OpenAI's GPT 3 [64], 3.5, and 4 [65] as well as emerging open-source initiatives such as BigScience's BLOOM [9a], Microsoft's Phi-2 [66] and Orca-2 [67], Mistral AI's Mistral 7B [68], and TII's Falcon-180B [69]. In January 2023, OpenAI's ChatGPT (based upon GPT-3.5) became the fastest-growing application in history when it reached more than 100 million monthly active users within two months of its introduction [70]. The subsequent media coverage surrounding these developments [71–73] further intensified public interest and concern.

> Accountability for AI systems and their impacts requires transparency around their design and creation and how they are intended to be used

While often categorized as technical, AI systems and their underlying data and models are *sociotechnical*. In other words, they combine the technical infrastructure and design with the social context in which they are designed, developed, evaluated, and deployed. Accountability for these systems and their impacts requires transparency around their design and creation and how they are intended to be used [74]. In recent years, alongside the exponential increase in data collection and the efforts to develop increasingly powerful machine learning models, there have been notable efforts calling attention to the need for documentation to accompany datasets, models, and AI systems, and to account for the process of creating them. These include Datasheets for Datasets [28], Dataset Nutrition Labels [29–31], Data Cards [32], Data Statements for Natural Language Processing [33–35], the Aether Data Documentation Template [36], Traditional Knowledge Labels [37], Nutritional Labels for Data and Models [42–44], Model Cards [45,46], Method Cards [47], FactSheets [48–54], and System Cards [55], among others.

These documentation efforts are complementary to Impact Assessments for AI systems, sometimes referred to as Algorithmic Impact Assessments, which are impact and risk evaluations carried out internally by the product team prior to development. These are usually intended to assess and categorize risks of the system, or to facilitate reflection on design choices. Impact assessments can result in documentation assets like those mentioned above; we discuss this further in Section III below. Both documentation and impact assessments are inspired, in part, by standardized approaches prevalent across diverse industries such as electronics, food and drug manufacturing, transportation, and civil and environmental engineering, that motivate and enforce product quality.

The public's recent concern has renewed regulatory focus on standards and requirements for machine learning models and systems. There are a number of proposed or passed legislative measures that suggest a move toward documentation requirements as an element of broader AI governance. These include the US White House AI Bill of Rights [18], EU AI Act [75], Canada's Artificial Intelligence and Data Act (AIDA) [76], Singapore's Proposed Advisory Guidelines on Use of Personal Data in AI Recommendation and Decision Systems (updated for PDPA) [77], and Brazil's draft AI legislation [78]. Notably, in May 2023, the U.S. congressional hearing "Oversight of A.I.: Rules for Artificial Intelligence" led by Senator Richard Blumenthal, brought further public attention to the need for AI documentation, including several calls for "fact sheets" or "nutrition labels" for datasets, models, or AI systems [27]. The October 2023 Hiroshima "Code of Conduct," signed by the G7 countries, called for "transparency reports…informed by robust documentation processes" that will "contribute to increased accountability" [79]. This is echoed by the October 2023 Executive Order on Safe, Secure, and Trustworthy Artificial Intelligence issued by the White House, which calls to "develop standards, tools, and tests to help ensure that AI systems are safe, secure, and trustworthy" [80]. The EU AI Act recommends assessments of risk, impact, and data privacy that, while distinct from the documentation outlined in the CLeAR Framework, will certainly support transparency and accountability broadly [81].

The recommendations and insights in this paper are distilled from years of experience and iterative development of documentation practices. Many of the authors of this paper have themselves been deeply involved in the major documentation projects and related research efforts from the last decade [21,23,28–31,34,35,38,39,41–45,47,49–56,74,82–84] and in creating AI documentation in industry contexts.

Key Themes and Considerations for AI Documentation Practices

Documentation is worthwhile for various stakeholders. It improves the understanding of practitioners creating or building datasets, models, or AI systems, which opens up opportunities to reflect on implicit and explicit decisions, ultimately enhancing the reliability of the systems they create. For organizations, it enables knowledge transfer across silos and encourages responsible use. Further, it provides information to users and potentially affected communities that can be used to determine the appropriateness of an AI system or its underlying data or models, thus helping inform consumer choice, advocacy work, regulation development, and regulation enforcement. It also enables recourse in the event of harms caused by or inquiries into the AI system, and accountability regarding who might be held responsible for those harms (see examples in Table 2).

This section provides background and context for AI documentation, as well as lessons learned from our prior work implementing AI documentation in a variety of contexts.

Importance of documenting datasets, models, and AI systems

In software engineering, documentation helps to improve the usability of a codebase or software system. However, AI systems (and the machine learning models they are built on) require more than just properly functioning code to work well across different use cases. For instance, how well a model works in a given context is greatly influenced by elements like the quality and representativeness of the training data and the cultural contexts it represents, and how the model was designed, built, trained, and fine-tuned. For this reason, a system developed in one country or cultural context, or one domain or domain context, may not perform well in another. The choice of benchmarks and evaluation criteria that were used to determine whether a model or system was ready to be released are also important for contextualizing its utility with respect to a specific use case: each test provides information about how well the model or system might work in the specific use contexts that the test represents.[1] As such, it is essential to also document upstream building processes (including dataset curation and model training) and intended use case scenarios. This information can inform future downstream use as the AI system changes hands: for example, if a third party builds on top of an existing model or AI system.

Documentation discloses potential risks, and can provide relevant information for mitigating them. It also can provide valuable information for appeal and recourse when harms have occurred. This is especially critical for models and AI systems that have high-impact applications, such as those used in healthcare, child welfare, immigration, financial lending, and public benefit distribution, where the potential for harm—such as loss of wealth or even death—is significant. This issue is particularly acute for subpopulations historically subject to unfair or unjust treatment, as historic injustices are reflected in the data and reproduced by models and AI systems.[2]

Waiting until the end of the development cycle to inspect datasets and AI systems prevents timely adjustments, and will dramatically increase the cost to address issues.

Documentation should occur throughout the lifecycle

While documentation should be created *in parallel* with development, it is common today to produce documentation *after* development. The post hoc practices around AI documentation pose challenges to developing technology that effectively accounts for foreseeable use and foreseeable harms. Furthermore, relevant information might be lost between development and documentation when documentation is created after development [31]. As such, we encourage developers to integrate documentation throughout AI development and deployment:

Models and AI systems should be inspected for bias and potential harms at all stages of the AI lifecycle. Often, the sociotechnical impacts of a model or AI system are only evaluated or critiqued after development and sometimes only after deployment. Even in the case where harms have not yet occurred, waiting until the end of the development cycle to inspect datasets and AI systems prevents timely adjustments, and will dramatically increase the cost to address issues, in terms of resources, reputation, and time. Indeed, in the "launch early and often" era of software development, it often falls on researchers, the media, or the public to identify and highlight harms of AI systems after they have already been deployed. Harms identification and mitigation only commencing after launch is too late; it is imperative that machine learning developers begin reflecting on potential harms in early stages of a project—for example, by completing an impact assessment—and continue to audit systems for risks both during the development process, as well as afterwards.

[1] For example, standard methods of model performance testing using hold-out data might not be sufficient to ensure acceptable machine learning model behavior in real-world settings. Yet external validation of machine learning models and systems is of special importance given they tend to fail silently, meaning they appear to function "normally" even when applied to data or contexts with very different properties than their training data [85,86].

[2] These include, but are not limited to, Indigenous peoples, people of color, women, LGBTQ+ communities, people with disabilities, immigrant communities, low-income and economically disadvantaged groups, and religious minorities.

Data and datasets also have a lifecycle that should be inspected for risks (ranging from privacy concerns to stereotypes and discriminatory content to how and when the data was produced or collected) [87]. Dataset creation introduces bias at every stage, including in collection, cleaning, processing, and labeling. As such, it is crucial to be intentional about which biases are acceptable for a particular use case or in tandem with mitigation procedures, and which are not. Choices that are made—either implicitly or explicitly—during the creation of a dataset will impact its downstream use, and so it is crucial to be intentional about these choices (especially as biases that arise unintentionally are no less harmful than intended bias). We note that data documentation is especially pertinent, and challenging, for large language models, whose efficacy draws directly from the size and scale of their training datasets, most of which are poorly documented today, if documented at all.

Choices that are made—either implicitly or explicitly—during the creation of a dataset will impact its downstream use.

In addition to documenting *throughout* the development process, we also encourage developers to *document the development process itself* and policymakers to scope regulation that takes into account the context and processes which led to the development of the dataset, model, or system. This can start, for example, from reflecting on a system's potential biases and harms when conducting an impact assessment. Although the final model or AI system may be the component most directly experienced by a user (for example, interacting with a chatbot, GPS system, or online reservation system), the product lifecycle encompasses much more. For instance, it may include: identifying a motivating problem or use case; determining which model(s) to use; gathering data for training, finetuning, and evaluating models; testing and validation; deployment in production; monitoring and updating; and so on. These stages typically occur non-linearly and are naturally iterative, meaning that documentation of the development process must be regularly updated to capture the dynamic nature of development.

Expanding the focus of documentation to be context-aware

In addition to documenting the process itself, and doing so throughout the AI lifecycle, the documentation should also be context-aware. Context-aware documentation expands the focus beyond the technical specifications of a dataset, model, or system to include sociotechnical information and decisions made during the process of development, many of which are critical to the performance of the AI system but would not otherwise be visible to someone using the system. This could include information such as initial motivation for the dataset, model, or system, how decisions about data collection and cleaning were made (including which communities were involved and why), and what particular metrics motivated design choices in models and systems. Gathering information on the implications of such sociotechnical decisions usually requires the participation of subject matter experts. It can further require the consultation of affected communities or other actors who are traditionally not considered part of the technical development pipeline, and thus whose perspectives are not typically included in documentation processes that don't call for such information. For example, Local Contexts is an initiative that enables the co-production of a dataset's relevant cultural metadata with Indigenous communities [88,89]. Including the knowledge of the populations represented in a dataset will preserve knowledge and contextual significance and honor lived experience, while driving a valuable exchange of information that can help mitigate downstream harms.

Risk and impact assessments in the context of documentation

In the context of AI governance, risk assessments, impact assessments, and documentation more broadly are closely related concepts that can be overlapping and complementary. While risk assessments concern the identification, analysis, and evaluation of threats and vulnerabilities, impact assessments go further by considering and, in some cases, measuring, implications for individuals, communities, and their environments. Risk-based AI governance approaches, such as those in effect or soon to be in effect in Canada [90] and the EU [81], mandate assessments as a way to identify the risk class of a system, which then determines the rules and—in many cases—documentation requirements applicable to the system in question.

Documentation is a lever that enables accountability.

Impact assessments are also increasingly common in the private sector [91] and civil society [92] as sociotechnical mechanisms to drive better design decisions. Typically, impact assessments are conducted *prior* to development by and for internal development teams in order to assess potential impacts of the system. This is in contrast to other documentation efforts that are focused on other parts of the AI system lifecycle, for a diversity of audiences, and covering a wide variety of system components (such as datasets and models). That said, these efforts are complementary, as risk and impact assessments include many overlapping themes to the documentation efforts we have discussed in this report (and cited in Figure 1 in the appendix), and will be easier to conduct and likely more accurate on systems that have Comparable, Legible, Actionable, and Robust documentation from which to draw information. Other forms of documentation can be informed by and should be updated following findings from impact assessments.

The opportunity to drive behavior through documentation requirements

The intention and value of documentation can shift dramatically depending on what is shared, with which audiences, and when. For example, internal documentation can provide standardized structures and create cultural norms that facilitate knowledge transfer, scientific rigor, and information sharing through a common vocabulary and approach. Furthermore, sharing the results of evaluations, inspections, and audits with third parties or oversight organizations can drive knowledge sharing across industry, promote understanding and increased public accountability, as well as support justice, equity, diversity, and inclusion (JEDI) initiatives. Documentation is a lever that enables accountability—especially when accountability systems (which can range from societal expectations, to voluntary standards, to regulation and legislation) require documentation review by either internal or external evaluators or auditors. In cases where the documentation is missing, inaccurate, or misleading—or in which the documentation reveals issues with the dataset, model, or AI system itself—mitigation procedures can minimize harm while the issues are addressed. Depending on the use case, such mitigation procedures can include limiting an AI system's automated decision making authority in favor of human review and oversight, or taking it offline. Additionally, documentation can provide standardized approaches and templates for running inspections, which can include compiling results, performing audits and evaluations, and communicating these results to relevant stakeholders.

Documentation requires robust organizational support

Documentation requirements are unlikely to be successful without robust organizational support, including additional resources for technical and workflow processes. In practice, the success of AI documentation largely depends on integrating documentation into existing tools and workflows to the extent possible [41] and on modifying current organizational and process structures as needed. Useful documentation takes time and requires appropriate skills. Thus, it is crucial that organizations prioritize and incentivize this task to develop an organic documentation culture. Having buy-in at the executive level and aligning the goals of documentation with the organization's goals is a first critical step to building momentum around real change in an organization's approach to documentation. In parallel, there is a powerful and important opportunity now to leverage regulation and policy as a mechanism to drive useful documentation practices.

The CLeAR Documentation Framework

We now introduce the CLeAR Framework. This framework aims to provide practitioners and policymakers with concrete principles to guide the implementation of **C**omparable, **L**egible, **A**ctionable, and **R**obust documentation of datasets, models, and AI systems. It also offers guidance for navigating the tradeoffs that can arise in practice when implementing these principles. By tradeoffs, we mean that at times the CLeAR principles are in tension with one another. For example, making documentation **comparable** limits the amount of customization, which means that for the sake of comparability, it can be beneficial to exclude certain details. Likewise, sometimes **legibility** might be at the expense of being very technical, being **actionable** might mean being less comprehensive, and **robustness** may be at odds with cost effectiveness. Since principles are often compelling in theory but challenging to realize in practice, our hope is that by discussing such tradeoffs, CLeAR can serve as a realistic guide for creating documentation while considering and understanding some of the choices that will need to be made.

The CLeAR Documentation Framework is intended to provide a foundation for choosing or designing an approach to documentation that is most suitable for a particular dataset, model, or AI system. It builds on and is aligned with other principles-based frameworks for documentation [44,56]. [3] The suitability of a particular form of documentation depends on a number of factors, including the context of the dataset, model, or system being documented, and the potential uses, goals, and audiences for the documentation. There is no one-size-fits-all approach when it comes to documentation; rather, our goal is to provide useful guidance and a shared vocabulary to support decision making.

The CLeAR Documentation Principles

Dataset, model, and AI system documentation should be **C**omparable, **L**egible, **A**ctionable, and **R**obust. These four principles are introduced below and further discussed throughout the remainder of the paper.

[3] Notably, the CLeAR Principles are aligned with the properties laid out by Stoyanovich and Howe: The principle *Legible* loosely maps to the property *Comprehensible, Actionable* to the properties Consultative and Concrete, and the principle *Robust* to the properties *Computable, Composable,* and *Concomitant.*

COMPARABLE: Documentation can be compared. It has similar components to the documentation of other datasets, models, or systems to permit or suggest comparison. It enables comparison by following a discrete, well-defined format in process, content, and presentation.

Comparability facilitates understanding by familiarizing various stakeholders with a consistent process and format for documentation. For instance, it enables potential users of a dataset, model, or AI system to evaluate it and contrast it with others, making it easier to judge relative performance, suitability, or impact.

While comparability is important, it requires a certain level of consistency or standardization. We propose that there is a base set of information relevant to all datasets, models, and AI systems that should be documented. For example, this includes *intended use.* For datasets specifically, this includes *source* as well as *dataset license.* Wherever possible, the presentation of these details should be standardized in format and content to provide for basic comparisons.

Since principles are often compelling in theory but challenging to realize in practice, our hope is that by discussing such tradeoffs, CLeAR can serve as a realistic guide for creating documentation while considering and understanding some of the choices that will need to be made.

It is important to stress that given the variety of datasets, models, and AI systems, it is likely not feasible for documentation to be standardized across all contexts and domains. Rather, the content, specific use cases, and the audience for the documentation itself should shape the documentation type and format, thus driving comparability that is adjusted to the need. For example, image processing algorithms applied in healthcare to detect potential cancerous cells in radiology images should adhere to a different set of required content for comparability than AI systems used to detect email spam. Also, the metrics used to evaluate these two distinct AI systems would be different. The first would likely have a measure like sensitivity as a primary metric since the costs of missing cancerous cells (so-called false negatives) are very high; in the second example, a primary metric like accuracy may be best suited, as marking legitimate emails as spam (false positives) can be as bad as failing to filter out spam emails (false negatives). Another example could be that documentation for a speech dataset may need to include information about the dialects spoken in each clip, whereas different demographic information would be appropriate for a vision dataset. There may also be differences in what is documented within different organizations based on their existing policies and systems.

LEGIBLE: Documentation can be read and understood by its intended audience. It is clear and accessible to this audience.

Legibility enables comprehension, which in turn supports the process of decision-making about the associated dataset, model, or AI system.

A key motivation for documentation is to provide transparency into the underlying asset (i.e., dataset, model, or AI system) with the goal of better understanding it, both in terms of its technical architecture, as well as the sociotechnical elements of how it was created and intended to be used. To be effective, the documentation must be legible, providing clear, comprehensive, and accurate information that enables those affected by it, those who will use it, and those who might leverage it to examine and assess it. It is important to keep in mind that what counts as legible will likely be different for different audiences [41]. Also, as with all approaches to transparency, documentation should be evaluated to ensure it helps its intended audience meet their goals [74,83,93]. While previous work has often focused on approaches intended for general audiences [82], we argue that different audiences might benefit most from having different versions of documentation that is legible for them.

In the spirit of clarity and comprehensibility, legible documentation should openly state relevant knowledge gaps, uncertainties, ongoing developments, and areas needing improvement that the developers consider important to inform on intended use. It also should articulate the expectation of updates (such as developments and improvements not foreseeable at the time of documentation) to capture the dynamic nature of the development process.

ACTIONABLE: The documentation can be acted on. It has practical value and is useful to its intended audience.

Actionable documentation contains the appropriate level of granularity and detail to enable informed decision making for its intended audience.

To be actionable, documentation must contain practical information that can be effectively used by stakeholders, for example for users to audit a dataset or evaluate model outputs [47]. Different stakeholders may have different uses for the information provided. For instance, designers may use the information presented in a model card to eliminate risky design ideas, create conditional designs to mitigate harms, provide transparency to end users, and advocate for the user perspective within their team [84]. Different stakeholders may also benefit from different kinds of information, and thus documentation may require multiple levels of technical depth (from highly technical to more qualitative or legal in nature) to be actionable for a broad set of users. For example, a data scientist may require technical information like performance and evaluation metrics, while a compliance officer may want to see details on the license and provenance. On the other hand, an affected consumer or user may be primarily interested in the system's biases or the fairness of its predictions.

In addition to considering audience(s) when determining an approach, practitioners should consider the information design and hierarchy within the documentation. For example, actionable documentation might contain "at a glance" information for some users, with deeper dive information for others. The visual design of the documentation, including designing it in an organized format and making it accessible for quick and easy reference, will also be crucial to make it actionable and should be considered as part of any approach.

ROBUST: The documentation can be sustained over time. It is kept up to date.

Robust documentation has durability and longevity that will help maintain the trust of stakeholders.

Without organizational or systemic support, documentation often falls by the wayside. Robust documentation remains effective over time, which may require regular updates. Its durability is enabled by sufficient support, integration, and flexibility of documentation practices. Documentation should be thought of as an integral part of the development process for datasets, models, and AI systems and must be prioritized and embedded across existing processes and expectations. Documentation should allow for additions as well as modifications as needs evolve over time. Stewards and managers of documentation should build in a process for updating over time even as they formalize the operational and organizational processes around the creation and publishing of documentation.

Using the CLeAR Framework

We have created the CLeAR Framework to help guide conversations through a shared vocabulary and understanding of why and how to leverage AI documentation. However, we know that these conversations are neither new nor simple. To help describe the complexities and real challenges of implementation, we have gathered below a number of **tradeoffs** that occur for each principle. In the subsequent section, we include some **open research questions** (Table 1) that are critical to address through both research and trial-and-error implementation in order to support progress over time. Examples to help guide the implementation of the CLeAR Framework can be found in the appendix (Table 2).

Comparable vs. customized: Even though there is not a one-size-fits-all approach, and we anticipate different standards across different datasets, models, or systems to be documented, documentation should follow a certain schema to enable comparison where possible. This will necessarily limit customization. It is conceivable that for the sake of comparability, certain information would be excluded. Similarly, creating comparable documentation may require alignment, potentially sacrificing customization, on certain aspects of the documentation process, such as when documentation begins, the kinds of expertise included, and general approaches for maintaining accuracy. Comparable *processes* can help produce comparable *documentation* and will mitigate inconsistencies from following different approaches, but may come at the expense of flexibility or customization.

Comparable vs. aligned with existing tools and processes: In practice, for efforts aimed at approaching AI responsibly—including AI documentation—to succeed, it is beneficial to align them with existing tools and processes within an organization to the extent possible [41,94]. Since tools and processes vary between organizations—and even, in some cases, between different teams within the same organization—this may lead to discrepancies between the content and format of documentation produced.

Comparable vs. actionable and legible for diverse audiences: Different consumers and stakeholders of documentation require different sets of information, both in terms of content and format. It is challenging to optimize the value of legible and actionable documentation to best serve stakeholders (including both creators and users of documentation) with differing values, needs, and goals. In such cases, bespoke or customized documentation may be beneficial, but may result in non-standardized information being shared, which limits comparison with other models or datasets.

Legible vs. (overly or insufficiently) detailed: Creating legible documentation for a specific audience requires including the right level of detail. Including extensive technical details may result in documentation being less comprehensible to audiences without the required background and expertise. Likewise, audiences with technical expertise might find the lack of detail in documentation written for less- or non-technical audiences less helpful. As another example, it may be necessary to include more background information and foundational definitions when preparing legible documentation for regulators as opposed to one's own team or organization. This challenge can sometimes be mitigated by producing different versions of documentation (e.g., ranging from technical to less- or non-technical), and offering some guidance for which versions of the documentation are intended for which audiences.

Actionable and legible vs. comprehensive: Documentation should encompass the necessary and relevant details for the intended stakeholders and use cases of the associated dataset, model, or AI system. However, documentation is most useful when appropriately concise. Thus, there is tension between being comprehensive and being concise: to be too brief can lose important content, but if documentation is exceedingly dense or lengthy, it will be less actionable.

Actionable vs. protecting privacy, security, and intellectual property: The full disclosure of information regarding the inner workings of a dataset, model, or AI system can come at the expense of exposing sensitive information, including private user data, protected intellectual property, or trade secrets that provide the basis for an organization's operational sustainability. To enable a balance between these tradeoffs, documentation standards should allow the owners and developers of a documentation asset to redact both personally identifiable information to protect privacy, as well as essential information or design elements that would limit the competitiveness of an organization or fall within protected intellectual property or trade secrets, and limit the release of information that can lead to security breaches, misuse, or fraud. Regulations will ultimately determine what information is required and what can be redacted.

Robust vs. resource intensive: The time and resources required for robust documentation, including keeping it current and updated, must be balanced with demands for other project needs and developments, some of which may be considered higher priority than documentation. Insufficient resource allocation for documentation may result in inadequate and unmaintained documentation. Furthermore, there may be high upfront costs for establishing robust documentation systems as well as a need to build capacities and resources for updating the documentation after the dataset, model, or AI system is initially deployed. Practitioners are encouraged to establish standardized methods to prioritize across documentation needs, including legacy, current, and future datasets, models, and systems, while keeping in mind that documentation can also aid in the successful development of existing or new datasets, models, or systems.

Open Questions and Directions for Exploration

As we have emphasized throughout, AI documentation is a relatively new process and what works best in practice is still being explored. In Table 1, we provide a list of open questions and directions for exploration, broken down by the CLeAR Principles. While considering these questions may provide initial guidance for practitioners in choosing the right approach to documentation to fit their needs and context, we acknowledge that answering these questions may also involve additional research as well as experimentation and iteration. This list is not exhaustive; it is intended as a starting point for further exploration.

Table 1:
Open Research Questions

COMPARABLE	• How should you **balance comparable** (consistent format and content) and **actionable** (including all relevant and useful information) documentation? • How might **striving for consistency** across the documentation assets of multiple datasets, models, or AI systems **limit the comprehensiveness** of the documentation? • How might documentation procedures and content that are standardized enable **interoperability** and **translatability** between groups, disciplines, and organizations? • What are the tradeoffs created by **automating** aspects of documentation through the use of tools? For example, automation may reduce **customization,** but likely enables greater **comparability** and **scalability.** • What **collaborative practices** and **partnerships or working agreements** are integral to the creation of documentation standards and the subsequent documentation processes that emerge from those standards?
LEGIBLE	• What information should be included in the documentation and what **level of detail** should be provided? What level of **technical knowledge** should be included (written for a layperson audience or technical audience, for what purpose)? • How can documentation be designed to be **accessible** and **understandable** for both technical and non-technical users? • What tools or formats can be leveraged to enhance the **accessibility** and **searchability** of the documentation, ensuring that it can be effectively navigated? • How can documentation incorporate **practical examples, use cases, and visualizations** to demonstrate the application and impact of the dataset, model, or system? • How can documentation be designed to enable users to understand the potential risks and **benefits** of a model or system, or the **contents** and relative **"health"** of a dataset?

Table 1 (continued):
Open Research Questions

Actionable	• Who are the **consumers** of documentation, and what information will they need? How do needs vary by the level of **technical expertise** and area or **domain of interest?** • How can researchers, companies, and developers be **incentivized to document shortcomings** and knowledge gaps that are necessary to understand for the documentation to be actionable? How can the financial or legal risks of disclosing shortcomings be mitigated? • Datasets, models, or systems may have to align with existing ethical or legal standards or frameworks. How can documentation of **privacy, data protection, proprietary knowledge, fraud and misuse prevention,** and **compliance** information help with this mapping? • How can documentation capture **uncertainty** and **limitations** associated with the dataset, model, or system to **empower users** to make informed decisions?
Robust	• What mechanisms can ensure that documentation remains **updated** to reflect any changes to the dataset, model, or AI system? Relatedly, how can documentation **versioning, version control,** and **archiving** be effectively managed? • How can documentation be made an **integral component** of the development and evaluation process? How can **existing developer workflows** and **managerial expectations** be adjusted to incorporate documentation practices? • How can organizations foster a **culture that values robust documentation** for developing high quality datasets, models, and systems, and for supporting **knowledge transfer** across teams and organizational domains? • What mechanisms can be used to encourage collaboration and sustainable ownership models of documentation, to facilitate collective responsibility for its maintenance and improvement? • How can documentation systems be designed to **accommodate evolving needs of stakeholders** such as engineers, data scientists, regulators, company executives, and the general public, enabling documentation to remain both **accurate** and **useful** in varied (and changing) environments? • How can documentation be made **resilient to potential risks** (such as data breaches, infrastructure failures, or natural disasters) to ensure that **critical information remains protected** and recoverable?

Implementation guidance and tips for practitioners

Below we list concrete suggestions that will assist practitioners in putting documentation into practice.

1. Look at past work on AI documentation

Researchers and practitioners have proposed methods to document datasets, models, and AI systems including Datasheets for Datasets [28], Dataset Nutrition Labels [29–31], Data Cards [32], Data Statements for Natural Language Processing [33–35], the Aether Data Documentation Template [36], Traditional Knowledge Labels [37], Nutritional Labels for Data and Models [42–44], Model Cards [45,46], Method Cards [47], FactSheets [49–54], and System Cards [55], and others. As these documentation approaches capture various use cases and were designed with different applications in mind, understanding their methodologies and motivations will be of help even for use cases or scenarios that have not yet been captured. These existing documentation approaches also illustrate best practices, such as requiring consistent language and protocols for writing and updating the documentation, and including an author list for transparency into which perspectives or domain expertise has been included in the creation of the documentation itself. Of course, as there is no one-size-fits-all approach to documentation, existing approaches may need to be adapted to fit the needs of a particular project or organization. More information about current documentation approaches can be found in Figure 1 in the appendix, which is a summary of AI documentation tools taken from Hugging Face's Model Card Guidebook [82].

2. Be realistic

Given limited resources and capacities within organizations, the approach to documentation implementation can be very different depending on the maturity and size of both the team and the system, the potential risks and harms of the system, and the scale and number of models or AI systems to document. For example, AI systems currently under development present an opportunity to start documentation in parallel with system development. By contrast, large legacy AI systems, which often include a multitude of individual models, can be exceptionally challenging to document retroactively. Additionally, any AI documentation will likely need to be revised often as new information becomes available or as the dataset, model, or AI system is used for additional tasks [52]. Further, it is important to be realistic about competing priorities and the complex nature of change management within the development process itself.

3. Start early and revise often

Much of today's AI documentation is created *after* development, which is less operationally efficient than parallel documentation creation. It also misses the opportunity to address issues as they arise during development and often makes it more expensive to address them later on. Documentation after development also has a tendency to reduce the accuracy of the information [31]. Thus for any dataset, model, or AI system in development, documentation should be created in parallel with development, beginning with the design of the dataset, model, or system, and throughout its lifecycle.

Starting early enables the opportunity to learn from shortcomings and revise the dataset, model or system development. Creating and adapting documentation is a learning process. To ensure accuracy, documentation should be created alongside technical milestones and continuously updated as needed, rather than delaying revisions until after development is complete [49].

4. Consider audience

When implementing documentation, a key element to keep in mind is the audience who will interact most closely with the documentation [27, [52]. Documentation that is targeted and designed for a very technical audience of software engineers, data scientists, and machine learning engineers will have a significantly different level of technical depth compared to documentation that aims at enhancing the understanding of end users. Separate forms of documentation can be produced for different audiences, or documentation can be designed in a "layered" way, including both the technical details as well as summaries and additional contextual information aimed at non-technical or business users.

5. Document regardless of size or scale

Regardless of the size or the scope of a dataset, model, or AI system, documentation is important. Models and systems of all sizes (including comparably "small" or "simple" ones) can reproduce bias, have unintended consequences, and cause outsized harm; this includes AI models and systems as well as non-AI (rule-based) models and systems. The same is true for datasets, where even comparably "small" or "simple" datasets can be highly biased and cause harm. On the other hand, the difficulty of documenting very large datasets does not reduce the importance of doing so [3].

Conclusion

As AI systems continue to permeate our lives, the importance of documenting how these systems are designed, trained, and deployed only grows. The demonstrated consequences and risks associated with AI deployment highlight the need for a deeper understanding of, and subsequent regulatory approach toward, data and data-driven systems. We face a great responsibility to create guardrails and implement changes that will guide the development of AI while mitigating potential harms and acknowledging the sociotechnical nature of these systems: The contexts in which they are used can result in wildly different outcomes, and the potential harms might disproportionately affect those who are already disenfranchised and marginalized in society.

Dataset, model, and AI system documentation are straightforward mechanisms for transparency into AI systems. The CLeAR Framework enables a foundation for designing and implementing documentation, while considering tradeoffs and encouraging holistic thinking about documentation needs. Building on the collective experience and perspective of a team that has worked at the forefront of AI documentation across both industry and the research community, we developed implementation guidance for practitioners, as well as context that may be helpful for policymakers. Given the complexity of AI governance, from data collection to model deployment, our goal is to establish a framework that serves as a guide for the consideration of documentation needs and priorities across datasets, models, and AI systems. Our hope is that this framework will help practitioners to create and use documentation, and support policymakers to better understand the importance of documentation and tradeoffs that should be considered for area-specific documentation regulation. Only through collective efforts can we ensure that AI is created and deployed responsibly. ■

Appendix

Table 2:
Examples of how existing documentation approaches address selected tradeoffs

COMPARABLE	Tradeoff: **Comparable vs. Customized** Documentation Example: **Dataset Nutrition Labels & IBM FactSheets** Analogous to nutrition labels on food, Dataset Nutrition Labels provide standardized information about a dataset through the content (for comparison) and the design (for legibility). Some key information include use cases, information about the inclusion of human subject data, provenance of the data, etc. However, due to the standardized nature of the questions, not all the information included on the labels is relevant or critical for all types of data. Furthermore, the manual creation method makes it challenging to document rapidly changing datasets with this documentation format. On the other hand, the FactSheets format invites data owners to write custom questions and provides application support for bespoke approaches to documentation. This can support more customized documentation, but it does present a challenge in comparing it with the documentation of other AI datasets, models, or systems, since the format may not be entirely matching between them.
LEGIBLE	Tradeoff: **Legible vs. Comprehensive** Documentation Example: **Meta's System Cards, Datasheets for Datasets** The System Cards standard was developed and designed to explain the functioning of Meta's systems (e.g., the Instagram Feed) in a way that is easy to access and comprehend. This indicates that the intended audience includes consumers and policymakers, who can leverage this concise documentation to better understand how the system works from a high, abstract level. However, this documentation sacrifices technical depth for its legibility: it omits detailed information about the architecture, model training steps or usage (feature selection, runtime choices, etc.), all of which is necessary for a technical understanding. Datasheets for Datasets approaches documentation from a different perspective, providing a detailed list of questions about a dataset's entire pipeline from its initial motivation through to its distribution and maintenance, at varying levels of technical specificity. While the authors encourage adaptation of the question list based on domain, organizational workflows, and other context, completing the full set of provided questions may result in documentation that is legible to technical experts but too detailed for a lay audience, both because of its content as well as the length of the final documentation.

Table 2 (continued):
Examples of how existing documentation approaches address selected tradeoffs

ACTIONABLE	Tradeoff: **Actionable vs. Privacy-protecting** Documentation Example: **MMIP Database documentation (Sovereign Bodies Institute)** The Missing and Murdered Indigenous Persons (MMIP) database[95] from the Sovereign Bodies Institute (SBI) logs cases of missing and murdered indigenous people. Out of respect for the sensitive nature of the data and of the families, SBI provides lightweight documentation in the form of a data dictionary only, and restricts access to additional information about the dataset, including the dataset itself, pending review of the type for the data request, usage intent, and organization or person requesting the data. In this case, SBI has chosen to provide documentation as a proxy for the data itself, limiting transparency into the data pending rigorous review of the request, thus prioritizing privacy over actionability.
ROBUST	Tradeoff: **Robustness vs. Resource allocation** Documentation Example: **HuggingFace Model Cards** HuggingFace Model Cards are an emerging attempt at model documentation, currently quite heterogeneous but trending toward standardization on the HuggingFace platform [96]. However, the team acknowledges that robust documentation is dependent upon processes that are both comprehensive and frictionless; to this end, the team conducted internal research that found that the most useful sections to include about a model (bias, risks, limitations) are also the most challenging and time-intensive to write [97]. This insight highlights the need for more streamlined workflow enhancements to capture this information, and it helps illustrate the tradeoff between robustness and resource allocation.

Figure 1.
Summary of AI Documentation Tools adapted from the Hugging Face's Model Card Guidebook [82]

Stage of ML System Lifecycle	Tool	Brief Description
Data	Datasheets (Gebru et al., 2018)	"We recommend that every dataset be accompanied with a datasheet documenting its motivation, creation, composition, intended uses, distribution, maintenance, and other information."
Data	Data Statements (Bender & Friedman, 2018) (Bender et al., 2021)	"A data statement is a characterization of a dataset that provides context to allow developers and users to better understand how experimental results might generalize, how software might be appropriately deployed, and what biases might be reflected in systems built on the software."
Data	Dataset Nutrition Labels (Holland et al., 2018)	"The Dataset Nutrition Label…is a diagnostic framework that lowers the barrier to standardized data analysis by providing a distilled yet comprehensive overview of dataset "ingredients" before AI model development."
Data	Data Cards for NLP (McMillan-Major et al., 2021)	"We present two case studies of creating documentation templates and guides in natural language processing (NLP): the Hugging Face (HF) dataset hub[^1] and the benchmark for Generation and its Evaluation and Metrics (GEM). We use the term data card to refer to documentation for datasets in both cases.
Data	Dataset Development Lifecycle Documentation Framework (Hutchinson et al., 2021)	"We introduce a rigorous framework for dataset development transparency that supports decision-making and accountability. The framework uses the cyclical, infrastructural and engineering nature of dataset development to draw on best practices from the software development lifecycle."
Data	Data Cards (Pushkarna et al., 2021)	"Data Cards are structured summaries of essential facts about various aspects of ML datasets needed by stakeholders across a dataset's lifecycle for responsible AI development. These summaries provide explanations of processes and rationales that shape the data and consequently the models."
Data	CrowdWorkSheets (Díaz et al., 2022)	"We introduce a novel framework, CrowdWorkSheets, for dataset developers to facilitate transparent documentation of key decisions points at various stages of the data annotation pipeline: task formulation, selection of annotators, plat- form and infrastructure choices, dataset analysis and evaluation, and dataset release and maintenance."
Models and Methods	Model Cards Mitchell et al. (2018)	"Model cards are short documents accompanying trained machine learning models that provide benchmarked evaluation in a variety of conditions…that are relevant to the intended application domains. Model cards also disclose the context in which models are intended to be used, details of the performance evaluation procedures, and other relevant information."

Figure 1. (continued)
Summary of AI Documentation Tools adapted from the Hugging Face's Model Card Guidebook [82]

Stage of ML System Lifecycle	Tool	Brief Description
Models and Methods	Value Cards Shen et al. (2021)	"We present Value Cards, a deliberation-driven toolkit for bringing computer science students and practitioners the awareness of the social impacts of machine learning-based decision making systems… .Value Cards encourages the investigations and debates towards different ML performance metrics and their potential trade-offs."
Models and Methods	Method Cards Adkins et al. (2022)	"We propose method cards to guide ML engineers through the process of model development…The information comprises both prescriptive and descriptive elements, putting the main focus on ensuring that ML engineers are able to use these methods properly."
Models and Methods	Consumer Labels for ML Models Seifert et al. (2019)	"We propose to issue consumer labels for trained and published ML models. These labels primarily target machine learning lay persons, such as the operators of an ML system, the executors of decisions, and the decision subjects themselves"
Systems	Factsheets Hind et al. (2018)	"A FactSheet will contain sections on all relevant attributes of an AI service, such as intended use, performance, safety, and security. Performance will include appropriate accuracy or risk measures along with timing information."
Systems	System Cards Procope et al. (2022)	"System Cards aims to increase the transparency of ML systems by providing stakeholders with an overview of different components of an ML system, how these components interact, and how different pieces of data and protected information are used by the system."
Systems	Reward Reports for RL Gilbert et al. (2022)	"We sketch a framework for documenting deployed learning systems, which we call Reward Reports…We outline Reward Reports as living documents that track updates to design choices and assumptions behind what a particular automated system is optimizing for. They are intended to track dynamic phenomena arising from system deployment, rather than merely static properties of models or data."
Systems	Robustness Gym Goel et al. (2021)	"We identify challenges with evaluating NLP systems and propose a solution in the form of Robustness Gym (RG), a simple and extensible evaluation toolkit that unifies 4 standard evaluation paradigms: subpopulations, transformations, evaluation sets, and adversarial attacks."
Systems	ABOUT ML Raji and Yang, (2019)	"ABOUT ML (Annotation and Benchmarking on Understanding and Transparency of Machine Learning Lifecycles) is a multi-year, multi-stakeholder initiative led by PAI. This initiative aims to bring together a diverse range of perspectives to develop, test, and implement machine learning system documentation practices at scale."

Bibliography

1. Rosenblatt F. The perceptron: a probabilistic model for information storage and organization in the brain. Psychol Rev. 1958;65: 386–408.
2. McCulloch WS, Pitts W. A logical calculus of the ideas immanent in nervous activity. Bull Math Biophys. 1943;5: 115–133.
3. Bender EM, Gebru T, McMillan-Major A, Shmitchell S. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?🦜. Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency. New York, NY, USA: Association for Computing Machinery; 2021. pp. 610–623.
4. Obermeyer Z, Powers B, Vogeli C, Mullainathan S. Dissecting racial bias in an algorithm used to manage the health of populations. Science. 2019;366: 447–453.
5. Norori N, Hu Q, Aellen FM, Faraci FD, Tzovara A. Addressing bias in big data and AI for health care: A call for open science. Patterns (N Y). 2021;2: 100347.
6. Incident 92: Apple Card's credit assessment algorithm allegedly discriminated against women. [cited 16 Mar 2024]. Available: https://incidentdatabase.ai/cite/92/
7. Incident 502: Pennsylvania county's Family Screening Tool allegedly exhibited discriminatory effects. [cited 16 Mar 2024]. Available: https://incidentdatabase.ai/cite/502/
8. Samoilenko SA, Suvorova I. Artificial Intelligence and Deepfakes in Strategic Deception Campaigns: The U.S. and Russian Experiences. In: Pashentsev E, editor. The Palgrave Handbook of Malicious Use of AI and Psychological Security. Cham: Springer International Publishing; 2023. pp. 507–529.
9. Mia Hoffmann HF. Adding Structure to AI Harm: An Introduction to CSET's AI Harm Framework. Center for Security and Emerging Technology; 2023 Jul pp. 10–29. Available: https://cset.georgetown.edu/wp-content/uploads/20230022-Adding-structure-to-AI-Harm-FINAL.pdf
10. Menz BD, Kuderer NM, Bacchi S, Modi ND, Chin-Yee B, Hu T, et al. Current safeguards, risk mitigation, and transparency measures of large language models against the generation of health disinformation: repeated cross sectional analysis. BMJ. 2024;384: e078538.
11. Zhou J, Zhang Y, Luo Q, Parker AG, De Choudhury M. Synthetic Lies: Understanding AI-Generated Misinformation and Evaluating Algorithmic and Human Solutions. Proceedings of the 2023 CHI Conference on Human Fac-

tors in Computing Systems. New York, NY, USA: Association for Computing Machinery; 2023. pp. 1–20.

12. Ziaei R, Schmidgall S. Language models are susceptible to incorrect patient self-diagnosis in medical applications. ArXiv. 2023;abs/2309.09362. doi:10.48550/arXiv.2309.09362
13. Simon J. Can AI be sued for defamation? In: Columbia Journalism Review [Internet]. 18 Mar 2024 [cited 7 Apr 2024]. Available: https://www.cjr.org/analysis/ai-sued-suit-defamation-libel-chatgpt-google-volokh.php
14. Nicholas Carlini G, Florian Tramèr, Stanford University, Eric Wallace UCB, Matthew Jagielski, Northeastern University, Ariel Herbert-Voss, OpenAI and Harvard University, Katherine Lee and Adam Roberts G, et al. Extracting Training Data from Large Language Models. USENIX. 2021. Available: https://www.usenix.org/system/files/sec21-carlini-extracting.pdf
15. A pro-innovation approach to AI regulation. In: GOV.UK [Internet]. [cited 24 Nov 2023]. Available: https://www.gov.uk/government/publications/ai-regulation-a-pro-innovation-approach/white-paper
16. Biden JR. Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence. 2023 [cited 24 Nov 2023]. Available: https://digitalcommons.unl.edu/scholcom/263/
17. Hiroshima process International Guiding Principles for advanced AI system. In: Shaping Europe's digital future [Internet]. [cited 24 Nov 2023]. Available: https://digital-strategy.ec.europa.eu/en/library/hiroshima-process-international-guiding-principles-advanced-ai-system?et_rid=928039398&et_cid=4962847
18. Blueprint for an AI bill of rights. In: The White House [Internet]. 4 Oct 2022 [cited 24 Nov 2023]. Available: https://www.whitehouse.gov/ostp/ai-bill-of-rights/
19. OECD legal instruments. [cited 7 Apr 2024]. Available: https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449
20. Taxonomy of Human Rights Risks Connected to Generative AI. United Nations Human Rights Office of the High Commissioner; Available: https://www.ohchr.org/sites/default/files/documents/issues/business/b-tech/taxonomy-GenAI-Human-Rights-Harms.pdf
21. Mitchell M. The pillars of a rights-based approach to AI development. In: Tech Policy Press [Internet]. 5 Dec 2023 [cited 7 Apr 2024]. Available: https://www.techpolicy.press/the-pillars-of-a-rightsbased-approach-to-ai-development/
22. JoAnn Stonier, Lauren Woodman, Majed Alshammari, Renée Cummings, Nighat Dad, Arti Garg, Alberto Giovanni Busetto, Katherine Hsiao, Maui Hudson, Parminder Jeet Singh, David Kanamugire, Astha Kapoor, Zheng Lei, Jacqueline Lu, Emna Mizouni, Angela Oduor Lungati, María Paz Canales Loebel, Arathi Sethumadhavan, Sarah Telford, Supheakmungkol Sarin, Kimmy Bettinger, Stephanie Teeuwen. Data Equity: Foundational Concepts for Generative AI. World Economic Forum; 2023 Oct. Available: https://www3.weforum.org/docs/WEF_Data_Equity_Concepts_Generative_AI_2023.pdf

23. Jagadish H, Stoyanovich J, Howe B. The Many Facets of Data Equity. J Data and Information Quality. 2023;14: 1–21.

24. AI Risk Management Framework | NIST. 2021 [cited 24 Nov 2023]. Available: https://www.nist.gov/itl/ai-risk-management-framework

25. Ehsan U, Saha K, De Choudhury M, Riedl MO. Charting the Sociotechnical Gap in Explainable AI: A Framework to Address the Gap in XAI. arXiv [cs.HC]. 2023. Available: http://arxiv.org/abs/2302.00799

26. Ehsan U, Singh R, Metcalf J, Riedl MO. The Algorithmic Imprint. arXiv [cs.CY]. 2022. Available: http://arxiv.org/abs/2206.03275

27. Oversight of A.i.: Rules for artificial intelligence. 16 May 2023 [cited 24 Nov 2023]. Available: https://www.judiciary.senate.gov/committee-activity/hearings/oversight-of-ai-rules-for-artificial-intelligence

28. Gebru T, Morgenstern J, Vecchione B, Vaughan JW, Wallach H, Iii HD, et al. Datasheets for datasets. Commun ACM. 2021;64: 86–92.

29. Holland S, Hosny A, Newman S, Joseph J, Chmielinski K. The Dataset Nutrition Label: A Framework To Drive Higher Data Quality Standards. arXiv [cs.DB]. 2018. Available: http://arxiv.org/abs/1805.03677

30. Chmielinski KS, Newman S, Taylor M, Joseph J, Thomas K, Yurkofsky J, et al. The Dataset Nutrition Label (2nd Gen): Leveraging Context to Mitigate Harms in Artificial Intelligence. arXiv [cs.LG]. 2022. Available: http://arxiv.org/abs/2201.03954

31. DNP Label Maker. [cited 24 Nov 2023]. Available: https://labelmaker.datanutrition.org/

32. Pushkarna M, Zaldivar A, Kjartansson O. Data Cards: Purposeful and Transparent Dataset Documentation for Responsible AI. Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency. New York, NY, USA: Association for Computing Machinery; 2022. pp. 1776–1826.

33. Bender EM, Friedman B. Data Statements for Natural Language Processing: Toward Mitigating System Bias and Enabling Better Science. Lee L, Johnson M, Toutanova K, Roark B, editors. Transactions of the Association for Computational Linguistics. 2018;6: 587–604.

34. McMillan-Major A, Bender EM, Friedman B. Data Statements: From Technical Concept to Community Practice. ACM J Responsib Comput. 2023. doi:10.1145/3594737

35. Bender EM, Friedman B, McMillan-Major A. A guide for writing data statements for natural language processing. Guide, Tech Policy Lab, Univ. Wash., Seattle. 2021. https://techpolicylab.uw.edu/wp-content/uploads/2021/10/Data_Statements_Guide_V2.pdf

36. Aether Data Documentation Template. In: Microsoft [Internet]. 25 Aug 2022. Available: https://www.microsoft.com/en-us/research/uploads/prod/2022/07/aether-datadoc-082522.pdf

37. TK labels. [cited 7 Apr 2024]. Available: https://localcontexts.org/labels/traditional-knowledge-labels/

38. Roman AC, Vaughan JW, See V, Ballard S, Torres J, Robinson C, et al. Open Datasheets: Machine-readable Documentation for Open Datasets and Responsible AI Assessments. arXiv [cs.LG]. 2023. Available: http://arxiv.org/abs/2312.06153

39. Data documentation. In: Microsoft Research [Internet]. 19 Feb 2021 [cited 7 Apr 2024]. Available: https://www.microsoft.com/en-us/research/project/datasheets-for-datasets/

40. Boyd KL. Datasheets for Datasets help ML Engineers Notice and Understand Ethical Issues in Training Data. Proc ACM Hum-Comput Interact. 2021;5: 1–27.

41. Heger AK, Marquis LB, Vorvoreanu M, Wallach H, Wortman Vaughan J. Understanding Machine Learning Practitioners' Data Documentation Perceptions, Needs, Challenges, and Desiderata. Proc ACM Hum-Comput Interact. 2022;6: 1–29.

42. Stoyanovich J. Revealing algorithmic rankers. In: Freedom to Tinker [Internet]. 5 Aug 2016 [cited 7 Apr 2024]. Available: https://freedom-to-tinker.com/2016/08/05/revealing-algorithmic-rankers/

43. Yang K, Stoyanovich J, Asudeh A, Howe B, Jagadish HV, Miklau G. A Nutritional Label for Rankings. Proceedings of the 2018 International Conference on Management of Data. New York, NY, USA: Association for Computing Machinery; 2018. pp. 1773–1776.

44. Stoyanovich J, Howe B. Nutritional Labels for Data and Models. IEEE Data Eng Bull. 2019;42: 13–23.

45. Mitchell M, Wu S, Zaldivar A, Barnes P, Vasserman L, Hutchinson B, et al. Model Cards for Model Reporting. Proceedings of the Conference on Fairness, Accountability, and Transparency. New York, NY, USA: Association for Computing Machinery; 2019. pp. 220–229.

46. Introducing the model card toolkit for easier model transparency reporting. [cited 24 Nov 2023]. Available: https://ai.googleblog.com/2020/07/introducing-model-card-toolkit-for.html

47. Adkins D, Alsallakh B, Cheema A, Kokhlikyan N, McReynolds E, Mishra P, et al. Method cards for prescriptive machine-learning transparency. Proceedings of the 1st International Conference on AI Engineering: Software Engineering for AI. New York, NY, USA: Association for Computing Machinery; 2022. pp. 90–100.

48. Hind M, Mehta S, Mojsilovíc A, Nair R, Ramamurthy KN, Olteanu A, et al. Increasing Trust in AI Services through Supplier's Declarations of Conformity. arXiv [csCY]. 2018. doi: https://doi.org/10.48550/arXiv.1808.07261

49. Hind M, Houde S, Martino J, Mojsilovic A, Piorkowski D, Richards J, et al. Experiences with Improving the Transparency of AI Models and Services. Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems. New York, NY, USA: Association for Computing Machinery; 2020. pp. 1–8.

50. Arnold M, Bellamy RKE, Hind M, Houde S, Mehta S, Mojsilović A, et al. FactSheets: Increasing trust in AI services through supplier's declarations of conformity. IBM J Res Dev. 01 July-Sep 2019;63: 6:1–6:13.

51. Arnold MR, Bellamy RKE, El Maghraoui K, Hind M, Houde S, Kannan K, et al. Generation and management of an artificial intelligence (AI) model documentation throughout its life cycle. US Patent. 11263188, 2022. Available: https://patentimages.storage.googleapis.com/fb/6e/76/cc3c09bfa940d9/US11263188.pdf

52. Richards J, Piorkowski D, Hind M, Houde S, Mojsilović A. A Methodology for Creating AI FactSheets. arXiv [cs.HC]. 2020. Available: http://arxiv.org/abs/2006.13796

53. AI FactSheets 360. [cited 24 Nov 2023]. Available: https://aifs360.res.ibm.com/governance

54. Piorkowski D, Richards J, Hind M. Evaluating a Methodology for Increasing AI Transparency: A Case Study. arXiv [cs.CY]. 2022. Available: http://arxiv.org/abs/2201.13224

55. Alsallakh B, Cheema A, Procope C, Adkins D, McReynolds E, Wang E, et al. System-Level Transparency of Machine Learning. Technical Report; 2022. Available: https://ai.meta.com/research/publications/system-level-transparency-of-machine-learning/

56. McMillan-Major A, Osei S, Rodriguez JD, Ammanamanchi PS, Gehrmann S, Jernite Y. Reusable Templates and Guides For Documenting Datasets and Models for Natural Language Processing and Generation: A Case Study of the HuggingFace and GEM Data and Model Cards. arXiv [cs.DB]. 2021. Available: http://arxiv.org/abs/2108.07374

57. Buolamwini J, Gebru T. Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. In: Friedler SA, Wilson C, editors. Proceedings of the 1st Conference on Fairness, Accountability and Transparency. PMLR; 23--24 Feb 2018. pp. 77–91.

58. Angwin J, Larson J, Kirchner L, Mattu S. Machine bias. In: ProPublica [Internet]. 23 May 2016 [cited 24 Nov 2023]. Available: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

59. Thoppilan R, De Freitas D, Hall J, Shazeer N, Kulshreshtha A, Cheng H-T, et al. LaMDA: Language Models for Dialog Applications. arXiv [cs.CL]. 2022. Available: http://arxiv.org/abs/2201.08239

60. Gemini Team, Anil R, Borgeaud S, Alayrac J-B, Yu J, Soricut R, et al. Gemini: A Family of Highly Capable Multimodal Models. arXiv [cs.CL]. 2023. Available: http://arxiv.org/abs/2312.11805

61. Glaese A, McAleese N, Trębacz M, Aslanides J, Firoiu V, Ewalds T, et al. Improving alignment of dialogue agents via targeted human judgements. arXiv [cs.LG]. 2022. Available: http://arxiv.org/abs/2209.14375

62. Taylor R, Kardas M, Cucurull G, Scialom T, Hartshorn A, Saravia E, et al. Galactica: A Large Language Model for Science. arXiv [cs.CL]. 2022. Available: http://arxiv.org/abs/2211.09085

63. Touvron H, Martin L, Stone K, Albert P, Almahairi A, Babaei Y, et al. Llama 2: Open Foundation and Fine-Tuned Chat Models. arXiv [cs.CL]. 2023. Available: http://arxiv.org/abs/2307.09288

64. Brown T, Mann B, Ryder N, Subbiah M, Kaplan JD, Dhariwal P, et al. Language models are few-shot learners. Adv Neural Inf Process Syst. 2020;33: 1877–1901.

65. OpenAI. GPT-4 Technical Report. arXiv [cs.CL]. 2023. Available: http://arxiv.org/abs/2303.08774

66. Hughes A. Phi-2: The surprising power of small language models. In: Microsoft Research [Internet]. 12 Dec 2023 [cited 7 Apr 2024]. Available: https://www.microsoft.com/en-us/research/blog/phi-2-the-surprising-power-of-small-language-models/

67. Orca. In: Microsoft Research [Internet]. 21 Nov 2023 [cited 7 Apr 2024]. Available: https://www.microsoft.com/en-us/research/project/orca/

68. Mistral AI. Mistral 7B. 27 Sep 2023 [cited 24 Nov 2023]. Available: https://mistral.ai/news/announcing-mistral-7b/

69. tiiuae/falcon-7b · Hugging Face. [cited 24 Nov 2023]. Available: https://huggingface.co/tiiuae/falcon-7b

70. ChatGPT sets record for fastest-growing user base - analyst note. Reuters. 2 Feb 2023. Available: https://www.reuters.com/technology/chatgpt-sets-record-fastest-growing-user-base-analyst-note-2023-02-01/. Accessed 24 Nov 2023.

71. The Google engineer who thinks the company's AI has come to life. The Washington Post. 11 Jun 2022. Available: https://www.washingtonpost.com/technology/2022/06/11/google-ai-lamda-blake-lemoine/. Accessed 24 Nov 2023.

72. Roose K. A Conversation With Bing's Chatbot Left Me Deeply Unsettled. The New York Times. 16 Feb 2023. Available: https://www.nytimes.com/2023/02/16/technology/bing-chatbot-microsoft-chatgpt.html. Accessed 24 Nov 2023.

73. Artificial intelligence report. In: WSJ [Internet]. The Wall Street Journal; [cited 24 Nov 2023]. Available: http://www.wsj.com/news/collection/artificial-intelligence-report-e48d5827

74. Liao QV, Wortman Vaughan J. AI transparency in the age of LLMs: A human-centered research roadmap. Harvard Data Science Review. 2024. doi:10.1162/99608f92.8036d03b

75. 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. In: EUR-Lex [Internet]. 21 Apr 2021 [cited 7 Apr 2024]. Available: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A52021PC0206

76. An Act to enact the Consumer Privacy Protection Act, the Personal Information and Data Protection Tribunal Act and the Artificial Intelligence and Data Act and to make consequential and related amendments to other Acts. In: LEGISinfo - Parliament of Canada [Internet]. 22 Nov 2021 [cited 7 Apr 2024]. Available: https://www.parl.ca/legisinfo/en/bill/44-1/c-27

77. Proposed Advisory Guidelines on use of Personal Data in AI Recommendation and Decision Systems. Available: https://www.pdpc.gov.sg/-/media/Files/PDPC/PDF-Files/Legislation-and-Guidelines/Public-Consult-on-Proposed-

AG-on-Use-of-PD-in-AI-Recommendation-and-Systems-2023-07-18-Draft-Advisory-Guidelines.pdf

78. Bill No. 2338, of 2023, PL 2338/2023 - Senado Federal. [cited 16 Mar 2024]. Available: https://www25.senado.leg.br/web/atividade/materias/-/materia/157233

79. Hiroshima process International Code of Conduct for advanced AI systems. In: Shaping Europe's digital future [Internet]. [cited 24 Nov 2023]. Available: https://digital-strategy.ec.europa.eu/en/library/hiroshima-process-international-code-conduct-advanced-ai-systems

80. The White House. FACT SHEET: President Biden issues Executive Order on safe, secure, and trustworthy artificial intelligence. In: The White House [Internet]. 30 Oct 2023 [cited 30 Nov 2023]. Available: https://www.whitehouse.gov/briefing-room/statements-releases/2023/10/30/fact-sheet-president-biden-issues-executive-order-on-safe-secure-and-trustworthy-artificial-intelligence/

81. Impact Assessment of the Regulation on Artificial intelligence. In: Shaping Europe's digital future [Internet]. [cited 7 Apr 2024]. Available: https://digital-strategy.ec.europa.eu/en/library/impact-assessment-regulation-artificial-intelligence

82. Hugging Face Model Card Guidebook. [cited 7 Apr 2024]. Available: https://huggingface.co/docs/hub/en/model-card-guidebook

83. Jennifer Wortman Vaughan HW. A Human-Centered Agenda for Intelligible Machine Learning. In: Marcello Pelillo TS, editor. In Machines We Trust: Perspectives on Dependable AI. MIT Press; 2021.

84. Liao QV, Subramonyam H, Wang J, Wortman Vaughan J. Designerly Understanding: Information Needs for Model Transparency to Support Design Ideation for AI-Powered User Experience. Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems. New York, NY, USA: Association for Computing Machinery; 2023. pp. 1–21.

85. Rabanser S, Günnemann S, Lipton ZC. Failing loudly: An empirical study of methods for detecting dataset shift. Adv Neural Inf Process Syst. 2018; 1394–1406.

86. D'Amour A, Heller K, Moldovan D, Adlam B, Alipanahi B, Beutel A, et al. Underspecification Presents Challenges for Credibility in Modern Machine Learning. arXiv [cs.LG]. 2020. Available: http://arxiv.org/abs/2011.03395

87. Hutchinson B, Smart A, Hanna A, Denton E, Greer C, Kjartansson O, et al. Towards Accountability for Machine Learning Datasets: Practices from Software Engineering and Infrastructure. Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency. New York, NY, USA: Association for Computing Machinery; 2021. pp. 560–575.

88. Anderson JE, Christen K. "chuck a copyright on it": Dilemmas of digital return and the possibilities for Traditional Knowledge licenses and labels. Museum Anthropology Review. 2013;7: 105–126.

89. Montenegro M. Subverting the universality of metadata standards: The TK labels as a tool to promote Indigenous data sovereignty. Journal of Documentation. 2019;75: 731–749.

90. Treasury Board of Canada Secretariat. Algorithmic Impact Assessment Tool - Canada.ca. 22 Mar 2021 [cited 8 Apr 2024]. Available: https://www.canada.ca/en/government/system/digital-government/digital-government-innovations/responsible-use-ai/algorithmic-impact-assessment.html

91. Microsoft Responsible AI Impact Assessment Template. Microsoft; 2022 Jun. Available: https://blogs.microsoft.com/wp-content/uploads/prod/sites/5/2022/06/Microsoft-RAI-Impact-Assessment-Template.pdf

92. Algorithmic impact assessment in healthcare. [cited 7 Apr 2024]. Available: https://www.adalovelaceinstitute.org/project/algorithmic-impact-assessment-healthcare/

93. Vera Liao Q, Varshney KR. Human-Centered Explainable AI (XAI): From Algorithms to User Experiences. arXiv [cs.AI]. 2021. Available: http://arxiv.org/abs/2110.10790

94. Madaio MA, Stark L, Wortman Vaughan J, Wallach H. Co-Designing Checklists to Understand Organizational Challenges and Opportunities around Fairness in AI. Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems. New York, NY, USA: Association for Computing Machinery; 2020. pp. 1–14.

95. MMIP database. In: SBI [Internet]. [cited 24 Nov 2023]. Available: https://www.sovereign-bodies.org/mmiw-database

96. Hugging Face Annotated Model Card Template. [cited 7 Apr 2024]. Available: https://huggingface.co/docs/hub/en/model-card-annotated

97. Hugging Face User Studies. [cited 16 Mar 2024]. Available: https://huggingface.co/docs/hub/model-cards-user-studies

Acknowledgments

The authors would like to thank the following colleagues for theirthoughtful feedback on and contributions to this report: Emily Bender (University of Washington), Ashley Casovan (IAPP, Carleton University), Mahima Pushkarna (Google Research), and Kush R. Varshney (IBM Research).

We would additionally like to thank Lindsay Underwood, Alicia Pritt, and Laura Manley at the Shorenstein Center for their support publishing this paper.

Contact: Kasia Chmielinski and Sarah Newman