Annotation and Benchmarking on Understanding and Transparency of Machine learning Lifecycles (ABOUT ML)

Table of Contents

Chapter 2: Current Recommendations on Documentation for Transparency in the ML Lifecycle

2.1 Demand for Transparency in ML Systems

Transparency requires that the goals, origins, and form of a system be made clear and explicit to users, practitioners, and other impacted stakeholders seeking to understand the scope and limits of its appropriate use. This is especially challenging in the context of ML systems, where models can encode complex and unintuitive relationships between inputs and outputs, making it challenging to naturally infer the details of what guided the process leading to a particular outcome.

As a result, many organizations include transparency as a core value in their statements of AI principles. Such statements can be an important first step[1] toward making transparency a focus of everyone contributing to ML development. Declared AI principles of course need to be translated into organizational processes and practical requirements for product decisions in order to become actualized. That translation is hard work and relies on building an institution with the systems and processes that can enact the principles put forth.

For instance, AI principles released by Google, IBM, Microsoft, and IEEE all include transparency clauses.[2] IEEE specifies that having AI systems "operate in a transparent manner" was a main goal of their release of their "Ethically Aligned Design" recommendations. Microsoft also names transparency as a core value, saying "AI systems should be understandable." IBM centers their entire principles statement on "Trust and Transparency" and Google, although not explicitly using the term transparency, states that "We will design AI systems that provide appropriate opportunities for feedback, relevant explanations, and appeal. Our AI technologies will be subject to appropriate human direction and control."

Broader studies analyzing the full scope of AI principle statements including those from governments, NGOs, academia and industry further reveal a focus on transparency - of 50 AI

---

[1] Friedman, B, Kahn, Peter H., and Borning, A., (2008) Value sensitive design and information systems. In Kenneth Einar Himma and Herman T. Tavani (Eds.) *The Handbook of Information and Computer Ethics.*, (pp. 70-100) John Wiley & Sons, Inc. http://jgustilo.pbworks.com/f/the-handbook-of-information-and-computer-ethics.pdf#page=104; Davis, J., and P. Nathan, L. (2015). Value sensitive design: applications, adaptations, and critiques. *Handbook of Ethics, Values, and Technological Design: Sources, Theory, Values and Application Domains.* (pp. 11-40) DOI: 10.1007/978-94-007-6970-0_3. https://www.researchgate.net/publication/283744306_Value_Sensitive_Design_Applications_Adaptations_and_Critiques; Borning, A. and Muller, M. (2012). Next steps for value sensitive design. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (CHI '12). (pp 1125-1134) DOI: https://doi.org/10.1145/2207676.2208560 https://dl.acm.org/citation.cfm?id=2208560

[2] Pichai, S., (2018). AI at Google: our principles. *The Keyword.* https://www.blog.google/technology/ai/ai-principles/; IBM's Principles for Trust and Transparency. *IBM Policy.* https://www.ibm.com/blogs/policy/trust-principles/; Microsoft AI principles. *Microsoft.* https://www.microsoft.com/en-us/ai/our-approach-to-ai; Ethically Aligned Design – Version II. *IEEE.* https://standards.ieee.org/content/dam/ieee-standards/standards/web/documents/other/ead_v2.pdf

principle statements documented through the Linking AI Principles (LAIP) project,[3] 94% (47) explicitly mention transparency. Similarly, 87% and 88% of principle statements surveyed in two other concurrent studies[4][5] reference "transparency." In fact, transparency is often highlighted as the most frequently occurring principle in these survey studies, and has been named "the most prevalent principle in the current literature."[6]

However, the challenge of translating the high-level ethical ideal of transparency into concrete engineering processes and requirements has been repeatedly referenced as a major challenge in terms of making the assertion of the value's importance, and translating that ideal into real world impact.[7] Although study after study confirms that meaningful progress cannot be made until ethical ideals are operationalized,[8] the inconsistency with which high level principles such as transparency are interpreted across different contexts, organizations and even teams makes it difficult to design consistent practical interventions. The lack of practical theory around these ethical ideals also serves as a roadblock to facilitating outside auditing from interested parties looking to hold AI system developers accountable[9], and can impede or slow down the responsible deployment of these models.

2.2 Transparency Through Documentation

One simple and accessible approach to increasing transparency in ML lifecycles is through an improvement in both internal and external documentation norms. For an increasingly concerned public or auditing organization, externally distributed and thorough documentation on ML system components is essential to earning and maintaining trust, and minimizing the misuse of these systems. External documentation and reporting requirements can also provide teams with an argument for more resourcing to implement transparency processes. Internal documentation is also vital, serving to improve communication between collaborating teams.

---

[3] Zeng, Y., Lu, E., and Huangfu, C. (2018) Linking artificial intelligence principles. *CoRR* https://arxiv.org/abs/1812.04814.

[4] Jessica Fjeld, Hannah Hilligoss, Nele Achten, Maia Levy Daniel, Sally Kagay, and Joshua Feldman, (2018). Principled artificial intelligence - a map of ethical and rights based approaches, *Berkman Center for Internet and Society,* https://ai-hr.cyber.harvard.edu/primp-viz.html

[5] Jobin, A., Ienca, M., & Vayena, E. (2019). Artificial Intelligence: the global landscape of ethics guidelines. *arXiv preprint arXiv:1906.11668*. https://arxiv.org/pdf/1906.11668.pdf

[6] Jobin, A., Ienca, M., & Vayena, E. (2019). Artificial Intelligence: the global landscape of ethics guidelines. *arXiv preprint arXiv:1906.11668*. https://arxiv.org/pdf/1906.11668.pdf

[7] Whittlestone, J., Nyrup, R., Alexandrova, A., & Cave, S. (2019, January). The Role and Limits of Principles in AI Ethics: Towards a Focus on Tensions. In *Proceedings of the AAAI/ACM Conference on AI Ethics and Society, Honolulu, HI, USA* (pp. 27-28). http://www.aies-conference.com/wp-content/papers/main/AIES-19_paper_188.pdf; Mittelstadt, B. (2019). AI Ethics–Too Principled to Fail? https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3391293

[8] Greene, D., Hoffmann, A. L., & Stark, L. (2019, January). Better, nicer, clearer, fairer: A critical assessment of the movement for ethical artificial intelligence and machine learning. In *Proceedings of the 52nd Hawaii International Conference on System Sciences*. https://scholarspace.manoa.hawaii.edu/handle/10125/59651

[9] Raji, I. D., & Buolamwini, J. (2019). Actionable auditing: Investigating the impact of publicly naming biased performance results of commercial ai products. In *AAAI/ACM Conf. on AI Ethics and Society* (Vol. 1). https://www.media.mit.edu/publications/actionable-auditing-investigating-the-impact-of-publicly-naming-biased-performance-results-of-commercial-ai-products/

Internal documents also help build employee trust by outlining the nature of an individual or team's contribution to an overall system, giving opportunity for ethical objections and a more meaningful understanding of the extent of their personal participation in the creation of an end product. Beyond the artifact itself, however, the process of documentation itself is inherently valuable towards the goal of transparency, prompting critical thinking about the ethical implications at every step in the ML lifecycle and encouraging the steps required to understand and report a complete picture on system capabilities, limitations, and risks.

As a result, the overall process an organization needs to follow, and deep dives into the specific transparency documentation questions to address with regards to the ML system are currently among the more well-researched topics in the transparency space. Organizations (including many PAI Partners) have begun to implement recommendations from those publications, and such work is beginning to influence procurement processes and regulatory documentation requirements by governments.[10] PAI's ABOUT ML effort aims to synthesize this research and learnings from previous transparency initiatives into best practices and new norms for documentation on ML lifecycles.

Documentation for transparency is both an artifact (in this case, a document with details about the ML system, similar to a nutrition label on food) and a process (in this case, a series of steps people follow in order to create the document). Both of these interpretations are at the core of the initial effort of the ABOUT ML initiative, which focuses on developing documentation to clarify the details of specific ML systems, for the sake of improving the transparency of that system. Note that this differs from the goals of other documentation proposals, which aim to set legally binding restrictions,[11] set declarative value statements,[12] or propose guidelines on systems-level ethical restrictions of use in a more general sense, and beyond the scope of model development and deployment decisions.[13]

2.2.1 Documentation as a Process in the ML Lifecycle

Documentation for ML lifecycles is not simply about disclosing a list of characteristics about the data sets and mathematical models within an ML system, but rather an entire process that an organization needs to incorporate throughout the design, development, and

---

[10] Algorithmic Impact Assessment (2019) *Government of Canada* https://www.canada.ca/en/government/system/digital-government/modern-emerging-technologies/responsible-use-ai/algorithmic-impact-assessment.html

[11] Benjamin, M., Gagnon, P., Rostamzadeh, N., Pal, C., Bengio, Y., & Shee, A. (2019). Towards Standardization of Data Licenses: The Montreal Data License. *arXiv preprint arXiv:1903.12262*. https://arxiv.org/abs/1903.12262; Responsible AI Licenses v0.1. *RAIL: Responsible AI Licenses.* https://www.licenses.ai/ai-licenses

[12] *See Citation 5*

[13] Safe Face Pledge. https://www.safefacepledge.org/; Montreal Declaration on Responsible AI. *Universite de Montreal.* https://www.montrealdeclaration-responsibleai.com/; The Toronto Declaration: Protecting the right to equality and non-discrimination in machine learning systems. (2018). Amnesty International and Access Now. https://www.accessnow.org/cms/assets/uploads/2018/08/The-Toronto-Declaration_ENG_08-2018.pdf ; Dagsthul Declaration on the application of machine learning and artificial intelligence for social good. https://www.dagstuhl.de/fileadmin/redaktion/Programm/Seminar/19082/Declaration/Declaration.pdf

deployment of the ML system being considered.[14] Incorporating transparency documentation into the build process includes asking and answering questions about the impact of the ML system and provides an internal accountability mechanism via the documented answers which can be referenced at a later date. In order to be effective, this non-trivial process needs resourcing, executive sponsorship, and other forms of institutional support to become and remain a sustainable and integral part of every project.[15]

This documentation process begins in the ML system design and set up stage, including system framing and high-level objective design. This involves **contextualizing the motivation for system development** and articulating the goals of the system in which this system is deployed, as well as providing a clear statement of team priorities and objectives throughout the system design process. After documenting the context in which the system is being developed and why it exists, the next step occurs during planning for the ML development pipeline by creating **a detailed and well-documented overview of the components of the ML development pipeline** - from the data used to train the system to the specifics of the system architecture and output characteristics. After system development comes system deployment. Important considerations at this stage usually involve testing, and the exploration of various methods to evaluate a system's effectiveness in achieving the stated goals of the system, while minimizing any undesirable side effects in the outcome, so transparent documentation would note all of these steps. And finally, even after deployment, system maintenance and monitoring systems provide a method to ensure the continued functionality of the system and maintain quality in system performance, so these steps should be included in any documentation process. Throughout these stages and after deployment, system feedback is key for incorporating the perspectives of those most impacted by the system's deployment to ensure adherence to the system goals written down in the design and set up phase. Feedback also informs higher level iterations of the system objectives during the design phase. At the development phase, feedback can guide implementation. During deployment and afterwards, feedback continues to provide external information on whether the model system is operating as intended and can help teams to course correct as needed. Thus, how feedback was solicited and incorporated needs to be well documented throughout the ML lifecycle.

---

[14] Dobbe, R., Dean, S., Gilbert, T., & Kohli, N. (2018). A Broader View on Bias in Automated Decision-Making: Reflecting on Epistemology and Dynamics. https://arxiv.org/pdf/1807.00553.pdf

[15] Wagstaff, K. (2012). Machine learning that matters. https://arxiv.org/pdf/1206.4656.pdf ; Friedman, B., Kahn, P. H., Borning, A., & Huldtgren, A. (2013). Value sensitive design and information systems. In *Early engagement and new technologies: Opening up the laboratory* (pp. 55-95). Springer, Dordrecht. https://vsdesign.org/publications/pdf/non-scan-vsd-and-information-systems.pdf

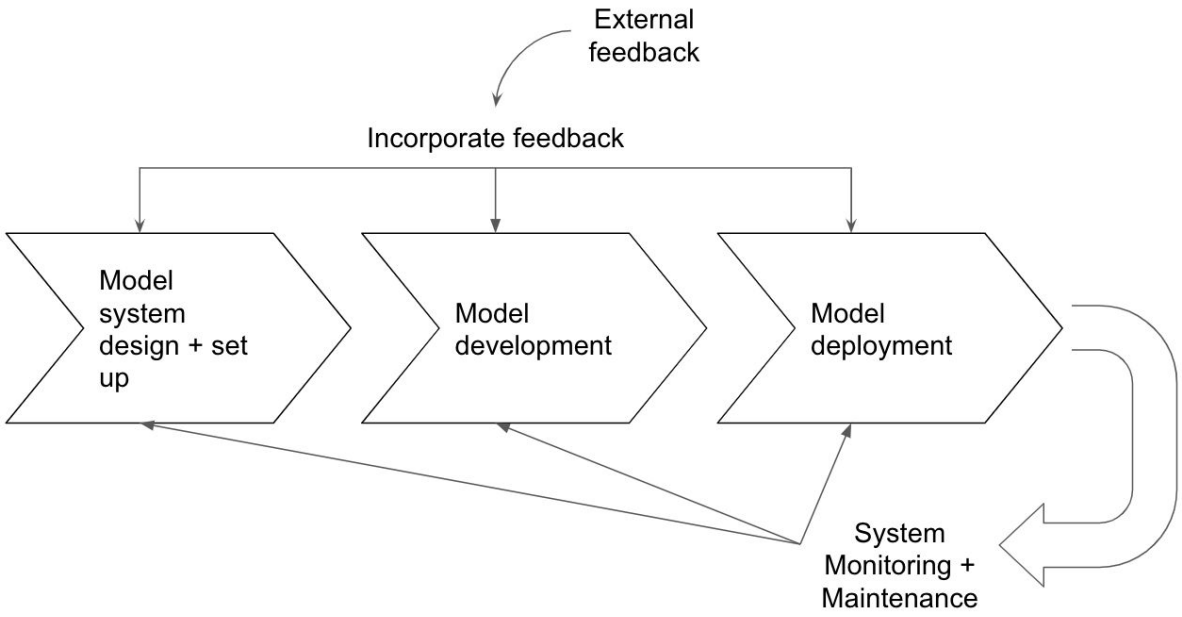

Figure 2.1 Overview of ML system lifecycle

2.2.2 Key Process Considerations for Documentation

At each step of this workflow, transparency and documentation need to be an explicit part of the discussion. What follows is a summary of the steps involved in the process and an overview of the challenges involved at each stage. As mentioned in the previous section, these steps are not necessarily sequential, and considerations at each step may come into play repeatedly and at various points in the ML system lifecycle.

1. System Feedback: Every step of the ML lifecycle needs to consider multiple perspectives, ideally through actively including individuals representing different stakeholders in the conversations. Key stakeholders are both internal (e.g., different departments, like Policy, Legal, Operations, and Sales) and external (e.g., civil society organizations, academia, policymakers, and people impacted by the technology). ML systems often operate in technology that impacts historically and presently marginalized communities, whose viewpoints are rarely considered by technology developers. A special effort should thus be made to consult those communities during each phase of model and system development in order to prevent rolling out a product with unanticipated adverse effects. Documenting how feedback was collected and incorporated in throughout this process may serve as a reminder to complete this important step.
2. System Design and Setup: Prior to developing a particular ML system, one must first design the context in which the model or models will be incorporated.[16] To

[16] Dobbe, R., Dean, S., Gilbert, T., & Kohli, N. (2018). A Broader View on Bias in Automated Decision-Making: Reflecting on Epistemology and Dynamics. https://arxiv.org/pdf/1807.00553.pdf

work towards incorporating transparency at this stage, teams and institutions must invest in the following:

a. Investing in Transparency as a Design Value: This begins with aligning the team and institutional setting to adequately resource the transparency process. Depending on the organization, this may require conversations and executive buy-in at multiple levels or a simple commitment and adding this as a performance metric or objective and key result (OKR). Before transparency can be achieved in a project, it must first be agreed upon as a prioritized design value for the team and organization involved.
b. Clarifying System Objectives: For this to be achieved, a clear-headed conversation about the ML product or system goals is required, including a detailed consideration of:
    i. how the specifications and requirements of the system serve the declared goals
    ii. what aspects are necessary versus optional
    iii. what risks or unacceptable use cases exist, and when might these be serious enough to halt the project
    iv. are there sensitive use cases and what limitations and oversight would be needed for those.

    Many teams are needed for this conversation, likely a combination of engineering, product, legal, business, operations, etc. In some cases, it is also important to bring in external perspectives like impacted parties to consult during this phase.

3. System Development: Now that the system requirements and goals have been specified, the questions for transparency move towards the tactical for both datasets, models and the overall systems. This stage involves identifying which datasets will be constructed, which systems will be built, and how all of them will be connected. There are often complications at the system level that arise from connecting complex datasets, models, and systems which impact additional stakeholders. This is the time to reflect about those and have a dedicated meeting together to identify what might go wrong and mitigation strategies to build into the system ahead of time (also known as a premortem).
    a. Data Documentation: As a driving factor of system development, custom or task-specific datasets are created, collected or developed for different uses. It is thus important to outline explicitly the intent and limitations behind the datasets involved in the testing and training of mathematical models within the system, as well as to disclose any other information relevant to the wider use of this dataset.
    b. Model Documentation: Model documentation involves a look at the characteristics of the intent behind the development of the model, and an account of decisions relevant to system reproducibility including algorithmic details and comments on the training process.
    c. System Documentation: System documentation discusses how datasets and mathematical models within the system are connected, what potential unforeseen complications may arise from those connections, how those complications are accounted for, and how the system will be used. This is the time to outline the ML requirements for the system and

discuss the alignment of these requirements to the stated goals and design values of the system.

4. System Deployment: Before releasing an ML system into the world, teams must ensure the system works as intended, consider risks from unintended uses - either via misuse or use in unintended contexts - and set forth a series of safeguards in place to minimize harm to those most affected by the ML system.
    a. System Testing: The selection of appropriate and fair evaluation criteria for measuring system behaviour and performance is an integral step in determining the effectiveness of an ML system. Making performance metrics and evaluation procedures clear and comprehensible will ensure an adequate understanding of the considerations involved before the decision to deploy the system or scale its use to a certain scope.
5. System Maintenance: After an ML system is released, it needs to be monitored and maintained to ensure expected performance and to check that all safeguards are working correctly. Documenting the planned and implemented monitoring process includes who does the monitoring and maintenance, what monitoring systems are in place, how they operate, and update and release timelines or trigger events.

Note that much of the process outlined above can take place before a single line of code is written. As the data is collected and the systems are built, there will likely be an iterative loop through some of these steps. For the best transparency results from documentation, answering an initial set of questions should take place before any product or feature is built. While certain detailed sections in the datasets and system phases may not be possible to answer before the datasets are collected or the systems are built, documentation must be built into the entire ML lifecycle from start to finish and cannot simply be added at the end.

2.3 Research Themes on Documentation for Transparency

There is substantial existing research on documentation for each of the steps outlined above. The following section provides a brief review of key insights from the current literature on 3 of the steps: System Design and Set up, System Development, and System Deployment.

2.3.1 System Design and Set up

Minimizing harm resulting from ML systems is a major theme in recent transparency research. Adverse impacts can result from intended and unintended misuse, which can result from applying an ML system in a context it was not designed for or using the system for a purpose it was not built for among other possibilities. Transparent documentation, especially about the system design and set up phase about how and why an ML system was built as well as inappropriate use contexts, reduces misuse by empowering builders, users, activists, policymakers and other stakeholders with the information necessary to call out intended and unintended misuse. Positive progress is happening through efforts such as the “Safe Face

Pledge"[17] and the "Montreal Declaration on Responsible AI,"[18] which improve the design and set up of an ML system by outlining dangerous use cases for the deployment of AI services in sensitive contexts and gaining public commitment against AI misuse from corporations through a signed pledge. Additionally, funding grants that promote the use of "AI for Good" provide incentives for positive use cases of AI, particularly to address the needs of traditionally underserved populations.[19]

Defining system feedback mechanisms from the outset is also essential for minimizing harm to the intended users and impacted non-users. Getting that feedback requires formalizing the inclusion of the perspectives of those most affected by the ML system, especially people from traditionally marginalized and underrepresented communities. Documenting feedback loops is a way to commit to the feedback process. The Diverse Voices method[20] from the Tech Policy Lab at the University of Washington is one way organizations can address this issue. The process involves identifying communities that will be highly impacted by the technology being considered, prioritizing based on which communities are least likely to be consulted by the developers of the technology, convening a panel of experiential experts from that community, asking for their feedback in a structured panel, incorporating that feedback into the design documents, and finally confirming with the panelists that their perspectives have been accurately reflected.

2.3.2 System Development

A central theme of promoting greater transparency in system development is detailed reporting on defining characteristics and intended uses of the system. There are well-researched sets of documentation questions meant to prompt thoughtful reflection prior to building datasets as well as models, including for different types of applications like NLP,[21] autonomous vehicles,[22] and other domains. These documentation templates are often modeled on those used in other industries, such as safety data sheets from the electronics industry[23] or nutrition labels from the food industry.[24] These suggested templates vary widely in

---

[17] Safe Face Pledge. https://www.safefacepledge.org/

[18] Montreal Declaration on Responsible AI. *Universite de Montreal.* https://www.montrealdeclaration-responsibleai.com/

[19] AI for Good Challenge. Google.org. https://ai.google/social-good/impact-challenge/; AI Impact Challenge. Microsoft. https://www.microsoft.com/en-us/ai/ai-idea-challenge

[20] Diverse Voices How To Guide. Tech Policy Lab, University of Washington. https://techpolicylab.uw.edu/project/diverse-voices/

[21] Bender, E. M., & Friedman, B. (2018). Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics*, *6*, 587-604.

[22] Ethically Aligned Design – Version II. IEEE. https://standards.ieee.org/content/dam/ieee-standards/standards/web/documents/other/ead_v2.pdf

[23] Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Daumeé III, H., & Crawford, K. (2018). Datasheets for datasets. https://arxiv.org/abs/1803.09010 https://arxiv.org/abs/1803.09010; Hazard Communication Standard: Safety Data Sheets. *Occupational Safety and Health Administration, US Department of Labor.* https://www.osha.gov/Publications/OSHA3514.html

[24] Holland, S., Hosny, A., Newman, S., Joseph, J., & Chmielinski, K. (2018). The dataset nutrition label: A framework to drive higher data quality standards. https://arxiv.org/abs/1805.03677; Kelley, P. G., Bresee, J., Cranor, L. F., & Reeder, R. W. (2009). A nutrition label for privacy. In *Proceedings of the 5th*

length and appearance, ranging from a single concise page of succinct statements, symbols and visualizations to upwards of 10 pages of detailed prose and graphs. Whether the documentation is meant for internal or external consumption also impacts length and contents, as internal documentation can be more detailed and thus can be longer.

A common focus across data-related templates is on clarifying why the dataset is being created and explicitly stating its intended use and limitations. Documentation questions across papers also consistently address the risks that arise at various stages of data creation and distribution, with the goal of encouraging practitioners to reflect on ethical concerns at every stage preceding data use and release. Some templates focus more on address specific risks like privacy. One goal for these templates is to create greater accountability as the ML project proceeds because team can refer back to initial goals to ensure ongoing consistency with their declared intentions.

Model- and system-level documentation efforts have since emerged from this earlier work on data documentation, introducing questions more specific to overall model objectives. This includes commentary on design decisions, such as model architecture and reporting on fair performance metrics,[25] as well as general "purpose, performance, safety, security, and provenance information to be completed by AI service providers for examination by consumers."[26]

In addition to reporting for collaborative knowledge and potential auditing, recent work has also suggested extending the role of documentation towards a legally binding contract similar to open software licenses.[27] Documentation could become a mechanism for restricting use, particularly in high-risk or high-impact scenarios out of scope of the dataset's suitable context. Although initial steps have begun on studying potential regulation of models and automation software,[28] most existing efforts focus on the promotion of best practices for model development rather than legally binding documentation. These efforts focus on broad recommendations for best practices for responsible machine learning[29] and ethics[30] to guide ML practitioners on ethical considerations as they prepare the model for training and deployment. These guidelines also include procedural guidance and suggestions specific to a

---

Symposium on Usable Privacy and Security (p. 4). ACM. http://cups.cs.cmu.edu/soups/2009/proceedings/a4-kelley.pdf

[25] Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., ... & Gebru, T. (2019, January). Model cards for model reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency* (pp. 220-229). ACM. https://arxiv.org/abs/1810.03993

[26] Hind, M., Mehta, S., Mojsilovic, A., Nair, R., Ramamurthy, K. N., Olteanu, A., & Varshney, K. R. (2018). Increasing Trust in AI Services through Supplier's Declarations of Conformity. https://arxiv.org/abs/1808.07261

[27] Benjamin, M., Gagnon, P., Rostamzadeh, N., Pal, C., Bengio, Y., & Shee, A. (2019). Towards Standardization of Data Licenses: The Montreal Data License. https://arxiv.org/abs/1903.12262

[28] Cooper, D. M. (2013, April). A Licensing Approach to Regulation of Open Robotics. In Paper for presentation for We Robot: Getting down to business conference, Stanford Law School.

[29] Responsible AI Practices. Google AI. https://ai.google/education/responsible-ai-practices

[30] Everyday Ethics for Artificial Intelligence. (2019). *IBM.* https://www.ibm.com/watson/assets/duo/pdf/everydayethics.pdf

particular use cases of concern, specifically facial recognition[31] and chatbots.[32] 2.3.3 System Deployment

The goal of documentation for system deployment is to encourage teams to incorporate strategies to ensure the model achieves its stated objective while minimizing undesirable and unanticipated side effects.

Several tools can accomplish this goal by testing for fairness and transparency. One of the early toolkits, FairTest, originated as part of the Unwarranted Associations framework in order to provide a practical tool for testing for unwanted and biased influences in a machine learning model.[33] As fairness definitions and the field evolved, more tools appeared to test a model for its performance according to these metrics,[34] with many tools released as part of corporate or open-source toolkits from Accenture,[35] IBM,[36] Facebook,[37] Google,[38] and Microsoft.

Although each of these toolkits remain grounded in statistical fairness definitions, some toolkits also emphasize the need for the qualitative assessment of these models to move towards fair evaluation practice. For instance, the What If Tool from Google heavily emphasizes enabling data visualizations to guide the practitioner's judgment on data diversity, and the Accenture toolkit involves a survey of high-level as well as detailed questions to consider before model deployment.

2.4 Suggested Documentation Formatting

The following is a more detailed discussion of documentation recommendations for specific ML lifecycle stages. In v0, these recommendations come from current research literature. Future versions will incorporate more feedback informed by current practices and pilots.

---

[31] Federal Trade Commission. (2012). Best Practices for Common Uses of Facial Recognition Technologies (Staff Report). *Federal Trade Commission*, *30*. https://www.ftc.gov/sites/default/files/documents/reports/facing-facts-best-practices-common-uses-facial-recognition-technologies/121022facialtechrpt.pdf

[32] Microsoft (2018). Responsible bots: 10 guidelines for developers of conversational AI. https://www.microsoft.com/en-us/research/uploads/prod/2018/11/Bot_Guidelines_Nov_2018.pdf

[33] Tramer, F., Atlidakis, V., Geambasu, R., Hsu, D., Hubaux, J. P., Humbert, M., ... & Lin, H. (2017, April). FairTest: Discovering unwarranted associations in data-driven applications. In 2017 IEEE European Symposium on Security and Privacy (EuroS&P) (pp. 401-416). IEEE. https://github.com/columbia/fairtest, https://www.mhumbert.com/publications/eurosp17.pdf

[34] Julius Adebayo (2017). FairML: Auditing Black-Box Predictive Models. Fast Forward Labs whitepaper. https://github.com/adebayoj/fairml , https:/Introducing AI Fairness 360/blog.fastforwardlabs.com/2017/03/09/fairml-auditing-black-box-predictive-models.html

[35] Kishore Durg (2018). Testing AI: Teach and Test to raise responsible AI. *Accenture Technology Blog.* https://www.accenture.com/us-en/insights/technology/testing-AI

[36] Kush R. Varshney (2018). Introducing AI Fairness 360. *IBM Research Blog*. https://www.ibm.com/blogs/research/2018/09/ai-fairness-360/

[37] Dave Gershgorn (2018). Facebook says it has a tool to detect bias in its artificial intelligence. *Quartz*. https://qz.com/1268520/facebook-says-it-has-a-tool-to-detect-bias-in-its-artificial-intelligence/

[38] James Wexler. (2018) The What-If Tool: Code-Free Probing of Machine Learning Models. *Google AI Blog*. https://ai.googleblog.com/2018/09/the-what-if-tool-code-free-probing-of.html

2.4.1 Suggested Documentation Sections for Datasets

A dataset is any collection of information. For this section, we specifically consider data that interfaces with the machine learning model. There are several types of datasets involved in a machine learning development lifecycle, all of which are important to document to offer the most complete understanding of an ML system. Datasets are used in different portions of the ML lifecycle to train, test, and validate the model (see sidebar for more info). Documentation on these different datasets varies greatly because they are used to accomplish different goals and may originate from different sources.

Transparency efforts are often correlated with the level of human involvement in the dataset. While most data in the world today is related to humans at some level, the more human-centric the dataset, the more security and explanation is needed to address concerns about privacy and potential misuse of the information. Sometimes the human involvement is obvious, such as with Census data, but other times the line can be difficult to identify, like in the case of commercial flight arrival time data (people are on the flight and are affected by the flight's timeliness). Thus, it is important for dataset authors to think carefully about how humans might be involved in the data, and how to protect and use that data responsibly.

The recommended documentation sections for datasets are as follows and will be described in detail below:

- Data Specification
    - Motivation
- Data Curation
    - Composition
    - Collection
    - Data Cleaning
- Data Integration
    - Uses
    - Distribution
    - Maintenance

2.4.1.1 Data Specification

2.4.1.1.1 Motivation

Documenting the motivation for producing a dataset provides a lever for accountability as the project proceeds, so that stakeholders can go back to the original intention for why the dataset was created and make sure the current trajectory tracks with the original goal, or to adjust accordingly. Moreover, sharing the original motivation openly can reduce the risk that the dataset will be repurposed for inappropriate uses down the line.

Pros/Cons

Writing down the motivation captures the context for the data, which can help downstream users make better informed decisions about how to use it.

One potential risk that businesses may see with documenting the motivation behind a dataset would be revealing too much information about their strategic goals for making the dataset.

Sample Documentation Questions

- For what purpose was the dataset created? (Gebru et al 2018)
- Who created this dataset and on behalf of which entity? (Gebru et al 2018)
- Who funded the creation of the dataset? (Gebru et al 2018)
- Curation Rationale (Bender and Friedman 2018)

2.4.1.2 Data Curation

2.4.1.2.1 Collection

The process for how the data was collected should be well-documented for end users of the system as well as any collaborators contributing to the development of the overall ML system. When data is collected from human subjects, this should include information about the consent and notification process. For example, are the subjects aware of all the data being collected about them, and are they able to opt out?

In addition, potential issues of sampling bias should be noted. For example, studies have found that certain minority communities are disproportionately targeted by police for arrest[39] which means that they are in effect being over-sampled in the data. As a result, arrest data would over-represent these communities even if they have similar crime rates to other communities.

Pros/Cons

The benefits of these disclosures include helping users of the dataset assess potential issues of bias. In addition, greater transparency around the collection process and whether the proper consent was obtained can give people more assurance that their data privacy is respected. In addition, these disclosures can allow companies to indicate that they have complied with relevant data privacy laws. Companies that make this information available might see a reputational or competitive advantage, as consumers might prefer using products built with models where the underlying data collection process is known, as it provides some assurance of access to evidence to leverage in the case of needing to identify and legally persecute misused models or data. For a company, more detailed documentation could effectively act as a legal shield against thirst party misuse, demonstrating on their part due diligence with regards to clarifying intended context of use, and thus providing evidence for any violations of that declared intent. Finally, documenting the data collection process enhances replicability.

Some potential negative effects of such disclosures, however, include possible legal, privacy, and intellectual property concerns depending on the level of granularity of the disclosures and whether any questionable practices were used for data collection.

Sample Documentation Questions

- What mechanisms or procedures were used to collect the data? How were these mechanisms or procedures validated? (Gebru et al 2018)
- If the dataset is a sample from a larger set, what was the sampling strategy? (Gebru et al 2018)

[39] Lum, K., & Isaac, W. (2016). To predict and serve?. Significance, 13(5), 14-19.https://rss.onlinelibrary.wiley.com/doi/full/10.1111/j.1740-9713.2016.00960.x

- Who was involved in the data collection process and how were they compensated, if at all? (Gebru et al 2018)
- Over what timeframe was the data collected? (Gebru et al 2018)
- Both genre and topic influence the vocabulary and structural characteristics of texts (Biber, 1995), and should be specified. Think of the nature of data sources and how that may introduce some bias. (Bender and Friedman 2018)

2.4.1.2.2 Cleaning

Although data cleaning can seem like a very straight-forward task, there is significant potential for bias to enter the dataset. Given that data cleaning often involves some degree of human labeling, disclosures should include information about the demographics of the labelers to help users gauge the potential for biased labeling or blindspots. There is both bias from the labelers themselves and bias from the choice of labels to include. For example, if sex is considered a binary variable, non-binary individuals are effectively unidentifiable in the data. Defining the taxonomy for the data is thus an important step in establishing the ground truth. In addition, ensuring inter-rater reliability is one important step to addressing the potential for bias from human labelers.

Pros/Cons

Benefits of such disclosures include replicability and clarifying potential biases. Model developers using the dataset can better understand what they can or cannot do with the data and any leaps in logic are more apparent. The transparency created by these disclosures can also encourage data collectors to ensure that their data labeling practices align with the original purposes they envisioned for the data.

Some potential downsides include that there might be some privacy concerns for the labelers depending on how much information about them is disclosed. In addition, data cleaning and labeling can be a complex and multi-layered process, so accurately relaying the process can be difficult. Finally, there might be some concerns that making the process for labeling extremely transparent might make any thresholds in the labeling process less meaningful. For example, if it is revealed that the threshold between a high and low score on a desirable attribute is 60, individuals near the threshold might try to game the system in order to achieve a high score. This issue is an instance of the concept known as Goodhart's Law and reflects the fact that once a measurement becomes a target, it may no longer be a useful measurement.

Sample Questions

- Was any preprocessing/cleaning/labeling of the data done? (Gebru et al 2018)
- Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data (e.g., to support unanticipated future uses)? (Gebru et al 2018)
- Which data instances were filtered out of the raw dataset and why? What proportion of the "raw" dataset was filtered out during cleaning?
- What are the demographic characteristics of the annotators and annotation guideline developers? (Gebru et al 2018)

2.4.1.2.3 Composition

It is vital to make it clear to users what is in the dataset. This reflects the operationalization of the motivation section above. In addition to a list of the label annotations

and metadata in the dataset, it is important to include information about how representative the dataset is and the potential for sampling bias or other forms of bias. For example, in the context of natural language processing (NLP), it would be relevant to include information about whose worldview the data reflects, or the original language of the text.[40] In general, background on the demographics reflected in the data can be useful for users to assess potential bias issues.

Pros/Cons

The benefits of making composition clear are that users can know what to expect from the dataset and how models trained on the data might perform in different domains. In disclosing composition, it is important for developers to refer back to their motivations for creating the dataset to ensure that the composition appropriately reflects those objectives.

Depending on the granularity of the description of composition, privacy could be an issue. As a general rule, developers should distinguish between what information is appropriate to share to whom and be very careful of disclosing any metadata or labels that might make the dataset personally identifiable.

When including information on the demographic composition of the data, developers should keep in mind that demographic taxonomies are not well-defined. Developers can look to the fields of sociology and psychology for existing standards, but should be aware that some taxonomies still be problematic in context. For example, a binary gender classification might not be appropriate.[41] In addition, it might not make sense to apply the American racial construct in another context.

Finally, there are still open research questions around the definition of "representativeness" in datasets. Representativeness depends on the context of the specific systems where the data is being used. Documentation should assist users with determining what the appropriate contexts are for use of the particular dataset.

Sample Questions

- What data does each instance consist of? (Gebru et al 2018)
- Is there a label or target associated with each instance? (Gebru et al 2018)
- Are there recommended data splits (e.g., training, development/validation, testing)? (Gebru et al 2018)
- Are there any errors, sources of noise, or redundancies in the dataset? (Gebru et al 2018)
- Detail source, author contact information and version history (Holland et al 2019)
- Ground truth correlations: linear correlations between a chosen variable in the dataset and variables from other datasets considered to be "ground truth" (Holland et al 2019)

---

[40] Bender, E. M., & Friedman, B. (2018). Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics*, *6*, 587-604.

[41] Katta Spiel, Oliver L. Haimson, and Danielle Lottridge. (2019). How to do better with gender on surveys: a guide for HCI researchers. *Interactions.* 26, 4 (June 2019), 62-65. DOI: https://doi.org/10.1145/3338283

2.4.1.3 Data Integration

2.4.1.3.1 Use

Stating the intended uses of a dataset can be helpful for users to understand whether the dataset is appropriate for their projects. In particular, such disclosures should include information on how the composition of the dataset or way in which it was collected and cleaned might affect future uses. Links to existing literature that uses the dataset can be helpful for illustrative purposes. This section would also include information on uses to be avoided and explanations of possible adverse consequences that might result from inappropriate uses.

Pros/Cons

Some advantages to these disclosures are that information about the intent of the dataset helps give potential users greater context and minimizes potential misuse. It is important to clarify common misconceptions. For example, data collected to classify people's facial expressions (smiling, frowning, etc.) might not be appropriate to use to classify people's underlying moods (happiness, sadness, etc.). This information also would help hold users of the dataset accountable.

Some challenges to making these disclosures include that it is difficult to identify all potential malicious uses of a dataset. In addition, malicious actors might purposefully use dataset in improper ways. Cautious legal departments may also be concerned about the possibility for liability with disclosing appropriate and inappropriate uses of the data. This can be mitigated through consultation with legal counsel.

Sample Questions

- Is there a repository that links to any or all papers or systems that use the dataset? (Gebru et al 2018)
- Are there tasks for which the dataset should not be used? (Gebru et al 2018)

2.4.1.3.2 Distribution

Distribution disclosures should relay how the dataset's creators will distribute the data for use and update the data, either internally between segments of their company or publically. This makes it easier for people to find and use the data and clarifies the intended audience. Such disclosures should include information about the accessibility of the dataset and the intended audience. Licensing and the timing of licenses and consent are important considerations in this disclosure process. Depending on how broadly the dataset is distributed, consent might need to be obtained from the subjects of the data. The distribution process should involve appropriate steps to preserve the privacy of these subjects. For example, there might be a log-in needed to access the dataset, and dataset users might need to sign a contract stipulating the conditions of use. This would have the added benefit of ensuring that if someone withdrew their consent to have their data included in the dataset, their data could be deleted in one place. In addition, the dataset should be properly anonymized prior to distribution and safeguards should be put in place to prevent de-anonymization.

Sample Documentation Questions

- How will the dataset will be distributed (e.g., tarball on website, API, GitHub)? Does the dataset have a digital object identifier (DOI)? (Gebru et al 2018)

- Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)? (Gebru et al 2018)
- Do any export controls or other regulatory restrictions apply to the dataset or to individual instances? (Gebru et al 2018)
- What documentation and metadata will you be releasing with the model? (Gebru et al 2018)

2.4.1.4 Maintenance

Providing information about the maintenance of datasets is important for helping users know whether they are using the latest dataset and whether the dataset will be kept up to date. The default assumption is generally that datasets are not maintained. If the dataset is not maintained, however, there can be concerns about the interpretability and applicability of the dataset for new projects. Developers who are interested in using the data should be informed about these potential issues so that they can draw appropriate inferences.

Moreover, if the dataset is not maintained, it can be difficult for individuals to remove their data. This is especially an issue with criminal records, which should be expunged periodically depending on local law. For developers for the EU, this can also create complications with GDPR. To ensure that users are using the latest version of the dataset, measures can be taken to ensure that users cannot download past versions of a dataset or that the dataset has an expiration date after which it is unusable without being updated.

Pros/Cons

Some benefits include that users would be able to better understand why and how the dataset changed. Proper maintenance techniques also make it possible for individuals to remove their content if they want to.

Some disadvantages are that maintaining datasets can be time- and resource-intensive, and being explicit about the plans for maintenance does to some extent require the dataset developer to follow through. In addition, doing maintenance well can be difficult, as there are potential issues with versioning and shifts in technology.

Sample Documentation Questions

- Is there an erratum (list of mistakes)? (Gebru et al 2018)
- Will the dataset be updated? If so, please describe how often, by whom, and how updates will be communicated to users? (Gebru et al 2018)
- When will the dataset expire? Is there a set time limit after which the data should be considered obsolete?

2.4.2 Suggested Documentation Sections for Models

Machine learning models use statistical techniques to make predictions based on known inputs.[42] They are incorporated into many real-world systems and business processes where prediction and estimation is valuable.

[42] Momin M. Malik. (2019). Can algorithms themselves be biased? *Medium.* https://medium.com/berkman-klein-center/can-algorithms-themselves-be-biased-cffecbf2302c

Model transparency is important because ML models are used in making decisions, and a society can only be accountable and fair if the decision-making within it is understandable, accountable, and fair. Clarifying the basis of a recommendation helps in achieving these objectives. When people know that the models they are designing will be accountable and understandable, they have strong reasons to aim at fairness.

Model documentation becomes even more important as machine learning gets incorporated into systems making high stakes decisions. For example, some states in the US are implementing ML-based risk assessment tools in the criminal justice system, and many companies have thus far refused to disclose what factors go into those decisions on the grounds of protecting "trade secrets." From a societal perspective, it is important that any products with so much potential impact on individual well-being are accountable to the people they impact, so it is particularly untenable for these products to remain wholly "black boxes." Other high stakes applications of machine learning include models that determine the distribution of public benefits, models used in the healthcare industry that impact consumer premiums under risk-based payment models, or facial recognition models applied in arrests.

The documentation steps outlined in this section apply to models built on static data, which does not change after being recorded, using various methods including supervised learning, unsupervised learning, and reinforcement learning. At this time, the guidelines below are less applicable to models that use streaming data, such as online learning models, where the dataset or metrics are dynamically changing.

It is very important to tailor the documentation to meet the specific goal of disclosing model-related information, including considering the most relevant audiences for achieving that goal. If the key audience is end users of a consumer-facing product, the level of disclosure should be less technical to avoid overwhelming the users. In particular, companies should avoid making disclosures so complicated that they reach a similar status as Terms of Service, which unfortunately can be so cumbersome that they serve only to protect institutions rather than inform or help the users. Policymakers and advocacy groups can play a role in ensuring that transparency disclosures do not evolve in that direction. In contrast, if the largest audience for a set of ML documentation is other developers at the same company, the disclosures can be much more technical and detailed. Of course, various details differ depending on the audience and context of use - one of the goals of later establishing best practices is to outline the requirements and expectations for transparent documentation in various common scenarios. For example, a non-technical one-pager may be suitable for the average consumer but is insufficient as au auditable document for policymakers and advocacy groups in high-stakes contexts.

Internal disclosures can be helpful to allow developers from the same organization to learn from each other's work. That said, internal disclosures should be careful to avoid legitimizing or spreading bad practices. The company should work independently to set and enforce high standards for models by making sure to provide enough human and capital resources to support the integration of transparency practices.

A common theme throughout this section is the importance of ensuring that the model disclosures do not create security or IP risks. Depending on what information about the model is disclosed and whether the documentation is for internal vs. external consumption, there

might be concerns that hackers might use this information to attack the system more effectively or that the company's trade secret protections might be compromised.

Finally, developers should be wary of Goodhart's Law when making model-related disclosures. Goodhart's Law is the principle that once a measurement becomes a target, it is no longer a good measurement. In this context, the worry would be that disclosing the details of the model might incentivize individuals to game the system by adjusting their actions to achieve their desired outcome.

Another problematic unintended consequence would be if companies hid key information by disclosing a high volume of less crucial information, which highlights the importance of looking at ML documentation as a process to follow which aims to prompt deep reflection about the impact of products that include ML models where documentation artifacts are a byproduct, rather than documentation for the sake of being able to claim that documentation was created.

2.4.2.1 Model Specifications

This section assumes that the intention for building the model has been documented earlier in the process, including task and system specification. There are three subjects to consider in specification: 1) about building models, 2) about evaluating models, and 3) extra specifications for models used in high-stakes or high-risk scenarios.

Within building models, key questions to document include the choice of structure (e.g. features, architecture, pretrained embeddings and other complex inputs), choice of output structure, choice of loss function and regularization, where random seeds come from and where they are saved, hyperparameters, optimization algorithm, and generalizability measured by how much difference in test they expect to see.

For evaluating models, it is key to discuss what kind of tests the model developer does regarding output, how the developer plans to identify and mitigate sampling bias (e.g., using a second source of truth to mitigate selection bias via reweighting), and how to evaluate model performance on real-world data relative to test set (what is the threshold of acceptable and what kind of use cases should be disallowed based on results).

If the use case involves high stakes for affected parties, it is essential to ensure and document that the choice of output structure and loss function appropriately encode and convey uncertainty both about predictions and across possible system goals.[43]

Pros/Cons

The benefits of documenting these details of model specification include reproducibility, spotting potential failure modes, and helping people choose between models for different use cases. There are potential security risks with revealing certain types of information. Proactive communication and thoroughly explaining the severity of risk across the spectrum of documentation and sharing the risk mitigation plan may help to alleviate these concerns. The risk of revealing "trade secrets" applies more to black box models, as disclosing some of these specifications may make it easier for others to reverse-engineer the model and to thus obtain information that a company considers trade secret.

---

[43] Eckersley, P. (2018). Impossibility and Uncertainty Theorems in AI Value Alignment (or why your AGI should not have a utility function). arXiv preprint arXiv:1901.00064.

Sample Documentation Questions

- What is the intended use of the service output? (Arnold et al 2018)
    - Primary intended uses
    - Primary intended users
    - Out-of-scope and under-represented use cases
- What algorithms or techniques does this service implement? (Arnold et al 2018)
- Model Details. Basic information about the model. (Mitchell et al 2018)
    - Person or organization developing model and contact information
    - Model date
    - Model version
    - Model type
    - Information about training algorithms, parameters, fairness constraints or other applied approaches, and features
    - Paper or other resource for more information
    - Citation details
    - License

2.4.2.2 Model Training

The focus of this stage in the ML lifecycle is on sharing how the model was architected and trained, and the process that was used for debugging.

Choices of ML model architectures have numerous consequences that are relevant to downstream users, so it is essential to document those choices. Did the designers choose a random forest, recurrent network, convolutional network, and why? What was the capacity of the model, how does it line up with the dataset size, and what are the risks of overfitting? What was being optimized for, and what regularization terms and methods were used?

Some particular considerations may apply to architectures for models that will be used for high stakes purposes: the wrong choice of optimization function or prediction objective can create significant risks of unintended consequences in deployment. In general, sufficiently high stakes ML systems should produce outputs that are explicitly uncertain both about predictions [44] and across different competing specifications of the system's goals.[45]

A separate datasheet should be attached to all datasets used in this process, likely to include training data and validation data used while adjusting the model. If federated learning or other cryptographic privacy techniques are used in the model, the datasheet may need to be adapted accordingly. Key questions for the validation data include how closely the data match real-world distributions, whether relevant subpopulations are sufficiently represented in the data, and whether the validation set was a hold-out set or if there was an effort made to be more representative of the real-world data distribution. Additionally, documentation should note any preprocessing steps taken, such as calibration corrections.

Another option is to add a link to the source code, which is again more likely for academic and open-sourced models than for industry/commercial models. It is important to document the version of all libraries used, github links, machine types, and hyperparameters involved in training. This increases reproducibility and helps future users of the model debug in case of difficulty. For very large datasets, sharing information about the compute platform and

---

[44] Partnership on AI. Report on Algorithmic Risk Assessment Tools in the U.S. Criminal Justice System, Requirement 5.

[45] Eckersley, P. (2018). Impossibility and Uncertainty Theorems in AI Value Alignment (or why your AGI should not have a utility function). arXiv preprint arXiv:1901.00064.https://arxiv.org/abs/1901.00064,

rationale behind hardware choices also helps future researchers and model developers to contextualize the model. Lastly, it is highly valuable to disclose how long the model took to train and with what magnitude of compute resources, as this allows future researchers to understand what level of resourcing a similar model would require to build.

Pros/Cons

As mentioned above, much of the documentation in this section is for the purpose of allowing other parties to build similar models, increasing reproducibility. Information on compute and hardware resources used also gives researchers the ability to judge how accessible the model is.

Debugging is the other large benefit of such robust documentation. For example, if a model has 94% percent accuracy in training, but 87% in test, knowing the original settings allows evaluators to identify whether this difference in performance comes for different settings, or from other factors. Combining model documentation with datasheets for the training data gives evaluators information to rule out performance changes due to changes in data or parameters. The evaluators can be internal stakeholders from testing teams or external stakeholders, like customers who purchase the model for deployment in their business processes.

Finally, this information also builds trust between research labs, the general public, and policymakers, as each party gains insight into how otherwise "black box" models were constructed.

Documentation for models can be both highly technical and lengthy, which runs into readability risks. It is important to present the information in a reader-friendly manner to ensure the hard work of documentation yields the benefits outlined above, and to prevent creating burdensome documentation that pushes the work unnecessarily onto users and consumers of the model.

Sample Documentation Questions

- What Training Data is used? May not be possible to provide in practice. When possible, this section should mirror Evaluation Data. If such detail is not possible, minimal allowable information should be provided here, such as details of the distribution over various factors in the training datasets. (Mitchell et al 2018)
- What type of algorithm is used to train the model? What are the details of the algorithm's architecture? (eg. a ResNet neural net). Include a diagram if possible.

2.4.2.3 Evaluation

After a model is trained, it needs to be tested. The goal of documentation in this section is to help users of the model to understand how the model was checked for accuracy, bias, and other important properties.

The specific metrics that the model will be tested on depend on the particular use case, so it is helpful for this documentation to include examples of which metrics apply for which use case. Testing models should also take into account the broader system the model is embedded into, so it is valuable to test the downstream effects of the model. These anticipated downstream effects should be documented as a baseline to evaluate against on an ongoing basis.

Breaking down accuracy and other metrics by specific subpopulations can be useful for detecting possible biases and ethical issues. For example, if a model has much lower predictive

accuracy or much higher false positive rates for one subpopulation, it might be problematic to use the model to make decisions for those subpopulations. The documentation should thus reflect those limitations clearly. Such disclosures would help ensure that the model is not later used to make unfair decisions for those subpopulations and could also give the developer some liability cushion against third party misuse since they have clearly stated the limitations of their particular model. Part of the goal of this section is to convey to the user that whether AI works is not binary and instead is a nuanced question depending on the use case and metrics of success.

The documentation should also discuss how the model developers checked for overfitting. For example, how did they construct the test set? Did they draw new test sets each time they trained the model for cross-validation purposes? It is important for the documentation to be very specific and to explore potential shortcomings of the model and the test set. While the test set is meant to be a reflection of reality, data is never perfect. Failing to account for known imperfections creates a risk that this documentation could lead to overconfidence and misuse of the model.

Pros/Cons

Disclosing the details of evaluation puts the system at risk of being more easily being manipulated by malicious actors, presenting a security risk. Indeed, the issue of Goodhart's Law, as discussed previously, can be a concern if individuals deliberately change their behavior to try to change the model's outcomes. That said, some activist groups engage in hacking to expose problematic aspects of an ML system with the goal of protecting vulnerable groups, so the increased potential for hacking can either be a positive or negative characteristic depending on one's perspective.

Sample Documentation Questions

- Which datasets was the service tested on? (Arnold et al 2018)
- Describe the testing methodology. (Arnold et al 2018)
- Describe the test results. (Arnold et al 2018)
- Are you aware of possible examples of bias, ethical issues, or other safety risks as a result of using the service? (Arnold et al 2018)
- Are the service outputs explainable and/or interpretable? (Arnold et al 2018)
- Metrics. Metrics should be chosen to reflect potential real world impacts of the model. (Mitchell et al 2018)
    - Model performance measures
    - Decision thresholds
    - Variation approaches
- Evaluation Data. Details on the dataset(s) used for the quantitative analyses in the card. (Mitchell et al 2018)
    - Datasets
    - Motivation
    - Preprocessing

2.4.2.4 Model Integration

Even if each portion of the model is thoroughly tested, validated, and documented, there are additional documentation and evaluation needs that arise when connecting a model

into a broader ML system. Validating how the models interoperate is important because different models might not work well together. It is important to test for how errors from all the models interact to find the corner cases that could pose problems in production. Latency changes when connecting models into a system, and that changes usability and reliability for users as well. The logistics of how the model is run - on the cloud vs on local machines - is another important factor to document because it changes how to handle client data in the pipeline, and clients in sensitive industries especially will be interested in the details of how that is handled. The system level documentation should also include information about system logs, pre and post-processing steps, and a summary of all the surrounding product and software design choices in the system which are not ML/AL related, but do impact how the ML/AI models operate in the system. Finally, there should be careful commentary on the system's vulnerability to adversarial inputs and whether there are system-level mitigations that have been put into place. This information is obviously sensitive and should be carefully considered before release to minimize the risk of malicious use of this information.

Pros/Cons

System level information can be especially useful for consumers, to help them determine how the system can be used in their context. It can also help regulators examine otherwise black box systems for compliance on privacy and data storage regulations. Of course, this can also be seen as a liability for a company, but on the other hand, it can equally be a force for increasing compliance with regulations. An internal process that supports accurate documentation of the entire ML system will also mitigate the risk of inaccurate documentation creating legal liability. For example, if one team writes that the system does not log, but a different team upstream of them does, it would be inaccurate to publish that documentation. Fortunately, adhering to a rigorous internal auditing system before publishing documentation can catch this type of inaccuracy, will likely have positive externalities of creating more cohesion between teams and preventing other issues that arise when functions become too siloed inside an organization.

Sample Documentation Questions

- What is the expected performance on unseen data or data with different distributions? (Arnold et al 2018)
- Was the service checked for robustness against adversarial attacks? (Arnold et al 2018)
- Quantitative Analyses (Mitchell et al 2018)
    - Unitary results
    - Intersectional results
- Ethical Considerations (Mitchell et al 2018)

2.4.2.5 Maintenance

As data, techniques, and real-world needs change, most models must be updated to ensure continued usefulness. There are many parameters of a model update which would be useful to document, including how and why an update is triggered, whether old models or parameters can still be accessed, who owns maintenance and updating, and guidance on reasonable shelf life of the mode. The first question to answer is why the model is being updated - for underlying data changes, process changes, or to improve model performance. Most of the answers to the questions above flow from this first one.

Updates can be triggered automatically based on time or specific metrics, or be active decisions evaluated from time to time. It is important to be clear about this so that people using the model can plan their workflows accordingly.

When models are updated, it can be useful to maintain access to previous versions or parameters.

Along with information on who owns the maintenance and update process, it is useful to know what the fallback plan is in case of personnel turnover or organizational changes. Having this documentation well-known makes it easier to make sure that model maintenance does not fall through the cracks during transitions.

The shelf life of a model, similar to the update schedule itself, can be dependent on time only or a combination of other metrics or factors. For example, it could be that when the underlying data distribution looks x% different, then the model should be evaluated on y criteria for whether it is still valid.

Pros/Cons

Completing all these steps of documenting an update process can be incredibly valuable both for people who will use the model and for the people building the model. For people using the model, they get more reliability and can better plan around the model based on update processes. For the people building the model, planning ahead for the update process can encourage an intentional approach and schedule up front. It also may encourage more precaution if the model developers have to document changes later, in the tradeoff between speed of progress and a precautionary principle.

There may be liability or timeline risks with thoroughly documenting the update process if the team misses deadlines or being locked into deadlines. One way to mitigate these risks would be to build in criteria/metric gates for updates rather than a hard timeline. For example, the update process will begin on the earlier of when X metric reaches Y level and 1 year from publication date. Also building in sufficient time to do the update is important because some updates are simple but others can take a long time if they are larger updates.

Sample Documentation Questions

- Does the service implement and perform any bias detection and remediation? (Arnold et al 2018)
- When were the models last updated? (Arnold et al 2018)
-