System-Level Transparency of Machine Learning

CHAVEZ PROCOPE, Meta AI, USA
ADEEL CHEEMA, Meta AI, USA
DAVID ADKINS, Meta AI, USA
BILAL ALSALLAKH, Meta AI, USA
NEKESHA GREEN, Meta AI, USA
EMILY MCREYNOLDS, Meta AI, USA
GRACE PEHL, Meta AI, USA
ERIN WANG, Meta AI, USA
POLINA ZVYAGINA, Meta AI, USA

Specialized documentation techniques have been developed to communicate key facts about machine-learning (ML) systems and the datasets and models they rely on. Techniques such as Datasheets, FactSheets, and Model Cards began the journey towards model documentation that provides ML explainability and transparency. Our proposal, called 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.

CCS Concepts: • **Software and its engineering → Documentation**; • **Computing methodologies → Machine learning**.

Additional Key Words and Phrases: System Cards, Transparency, Explainability

1 INTRODUCTION

With the rapid adoption of machine learning (ML) in practical applications, ML-specific documentation has become central to their transparency and to the user experience of different stakeholders of ML. In contrast to traditional software systems, the documentation of ML-based systems is an emerging area that poses new challenges and technical requirements. A variety of initiatives have been undertaken over the past few years to address these challenges. These initiatives aim to provide systematic ways to document ML-based systems.

Previous work on documenting ML-based systems have focused on different components of these systems. For example, Model Cards [Mitchell et al. 2019] aim to communicate key facts about individual models such as their intended use, training and evaluation data, and relevant metrics. Likewise, Datasheets for Datasets [Gebru et al. 2021] focus on providing various details about the datasets used to develop ML models. While the above solutions strongly advance the transparency of ML-based systems, they might not always be suited to capture system-level information. The following are two examples for such situations:

- Complex systems, such as ML-based content ranking, involve various models and datasets used to train those model, in addition to non ML-based technologies. The documentation solution therefore needs to communicate how data is processed over multiple components, including human-in-the-loop ones and non-ML routines.
- Individual models often involve outputs of other models such as generic embeddings or baseline models used for transfer learning as inputs. Accordingly, the documentation solution needs to account for such dependencies and their implications.

IBM's FactSheets [Arnold et al. 2019] aim to act as declarations of conformity for AI services by providing relevant details to the consumers of these services. The authors argue how services can consist of multiple models, and how assessing the conformity analysis should consider how these models fit together to provide the service.

We present an integrated solution that aims to provide enhanced transparency into ML-based systems. Our solution is designed to communicate both an overview and detailed information about such a system, and caters both to expert stakeholders including model developers, reviewers, and other stakeholders such as individuals who want to learn more about the ML underpinning their experiences.

2 BACKGROUND AND MOTIVATION

Providing transparency into ML systems involves distinct challenges. Some of these challenges are specific to ML systems, such as model interpretability, while others apply to software systems generally.

2.1 Transparency Through Documentation

The ABOUT ML initiative [Raji and Yang 2019] advocates documentation as a practical intervention to provide clarity into decision making in ML systems for stakeholders. The authors explain the value of both external and internal documentation such as establishing trust and demonstrating fairness [Holstein et al. 2019]. Furthermore, the authors argue that documentation is both an artifact and a process, showing how developing the documentation fosters ML developers to think critically about every step in the ML lifecycle. The initiative is ongoing, led by the Partnership on AI consortium which continues to actively develop this process.

Documenting software systems is a long-standing goal of software engineering. A wide variety of tools, techniques, and standards have been developed over decades to make the documentation process effective and efficient. Some of the emerging documentation initiatives for ML-based systems have adapted existing methods while others necessarily take a novel approach. Datasheets [Gebru et al. 2021], and to an extent Model Cards [Mitchell et al. 2019], grew from experiences with hardware specification documentation. Dataset Nutrition Labels [Chmielinski et al. 2020; Holland et al. 2018], FactSheets [Arnold et al. 2019], and Data Statements [Bender and Friedman 2018] are also prominent examples of these initiatives. Hind et al [Hind et al. 2020] report on the experience of ML teams in using FactSheets, and provide recommendations for easing the collection and flexible presentation of ML facts to promote transparency.

2.2 Transparency Through Model Interpretability

Lipton [Lipton 2018] analyzes in detail the notion of transparency in interpretability. The author argues how transparency in this context is the opposite of "black-boxness" as it helps understanding the mechanism by which the model works. Such understanding can be at the level of the entire model, at the level of individual components such as parameters, and at the level of the training algorithm.

Weller [Weller 2019] identifies different use cases of transparency, and distinguishes between two types of interpretability solutions to support them: global (how an overall system works) and local (explaining a particular prediction). The author surveys available techniques for both types and argues how the transparency enabled by them can provide insights into important model characteristics such as robustness and fairness.

2.3 Why ML Systems?

There is currently no consensus on the definition of artificial intelligence however there is recognition of the need to identify functional definitions. Krafft [Krafft et al. 2020] highlight how "AI researchers favor definitions of AI that

emphasize technical functionality, policy-makers instead use definitions that compare systems to human thinking and behavior." Where defining AI is challenging due to the variety of perspectives and disciplines involved, machine learning has reached point closer to an accepted definition, "the use and development of computer systems that are able to learn and adapt without following explicit instructions, by using algorithms and statistical models to analyze and draw inferences from patterns in data." [Press 2022].

Current model documentation approaches, which are necessarily descriptive in nature are often question and answer lists, these may fall short of providing explanations into systems where a series of models creates signals into other models that ultimately provide a user experience. System-level transparency provides insight into how an ensemble of models work together to make systems such as those that rank and recommend.

3 SYSTEM CARDS

Inspired by the solutions mentioned in Section 2.1, we propose System Cards as a means for documenting ML-based systems. These cards aim to provide insight into the system architecture and help better explain how the system operates. While it shares several aspects with previous work, our solution focuses primarily on system-level information. A system card provides an overview of several ML models that comprise an ML system, as well as details about these components, and a walkthrough with an example input.

Overview First. The entry point in a System Card is a page that provides a brief description of the system and its purpose, along with information about its authors and current version. In addition, this page has a high-level diagram of system components that communicates how they are used to achieve the desired functionality. Various diagrams have been devised to provide an overview of complex systems, such as UML diagrams for object-oriented software systems [Medvidovic et al. 2002]. In contrast, ML systems lack standard diagrams, with existing ones such as TensorBoard [Wongsuphasawat et al. 2017] focusing mainly on individual ML models. Figure 1 shows an example of an overview of a fictional content ranking system. The system leverages multiple data modalities that are typically present in user-generated content, specifically text, image, and video modalities. These pieces of the input are processed by multiple components of the ML system, in order to compute the final ranking.

Interactive Exploration. The user can access details about various components in a system card by interacting with the overview page. In particular, the user can access model cards [Mitchell et al. 2019] of the ML models in the system in order to examine the details of these models. Besides model cards, the user can also access a summary of the overall system performance about non-ML components such as human reviews.

A Walkthrough with an Example Input. Demonstration has become essential for sharing and testing ML solutions [Abid et al. 2019]. To improve transparency in ML systems, a system card can include a step-by-step demonstration of how the system processes an actual input. This helps stakeholders better grasp how the system works and appreciate the challenges it tackles. Moreover, the system card can enable the user to modify the input in order to examine how the system's response changes accordingly.

4 CHALLENGES AND FUTURE WORK

Documentation solutions that target ML systems have only started to emerge in recent years. The main challenge to these solutions is to increase their adoption among practitioners, by maximizing the value they add and minimizing the

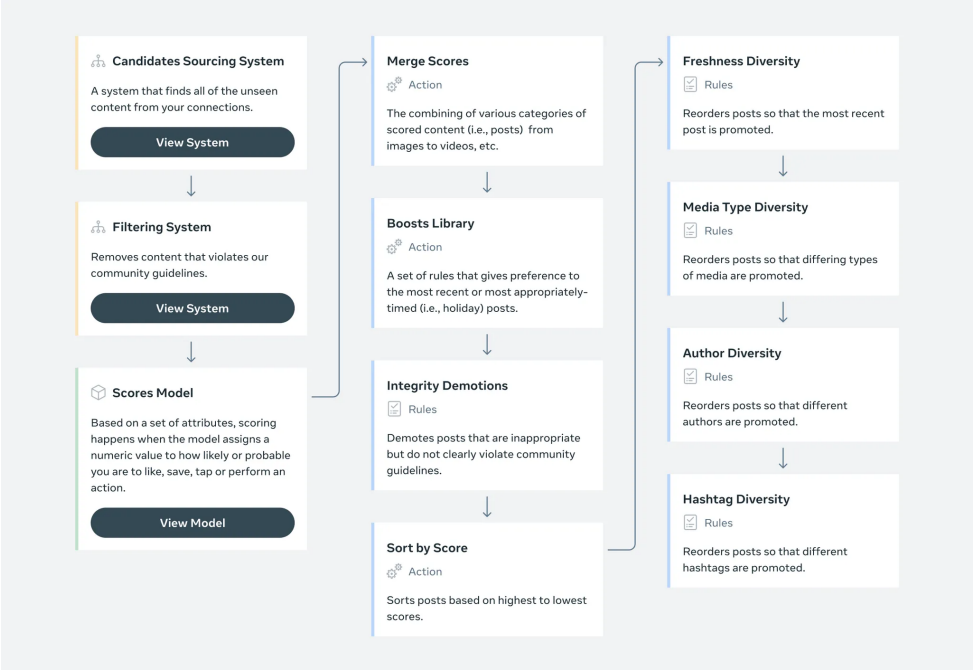

Fig. 1. An overview for an archetypal content ranking system showing how different components process various input modalities to compute the final ranking.

effort needed to create usable and valuable documentation artifacts. In the case of System Cards, the main challenges we identified are:

- **Automation:** The curation of System Cards currently largely depends on manual efforts. This includes creating the system diagram and the user interface that demonstrates how the system processes actual inputs. These artifacts require strong expertise in order to effectively simplify highly technical information. Automation has been shown to be useful in the creation of documentation artifacts such as Model Cards [Fang and Miao 2020]and FactSheets [Hind et al. 2020]. Likewise, solutions such as Gradio [Abid et al. 2019] can simplify the demonstration of ML systems. Inspired by these solutions, we are exploring avenues to reduce the manual effort involved in creating a system card.
- **Maintenance**: ML systems constantly evolve as their component models may often be retrained or replaced. Such updates would necessitate the System Cards to reflect these changes. Unlike in classical software systems, current ML documentation solutions rely largely on manual effort to maintain their content. We are exploring solutions to facilitate easier maintenance and provenance of System Cards in a highly agile ML landscape.
- **Security**: Providing information about how certain ML systems work could invite adversarial attacks, which could potentially risk harming the people who use or are affected by the system. Depending on the information provided in a System Card, malicious actors can use this knowledge to reverse-engineer the system. Therefore, it is important to strike the balance between ML transparency and security. We envision a solution in which different stakeholders have access to different levels of information about the system. For example, while some stakeholders may be more interested in assessing the appropriateness of the ML models used, end users may

be more interested in understanding which parts of their information are being used, for which purpose, and accordingly how they can opt-out of the provided services.

We plan to continue iterating on the system card concept as we learn from our end users and experts. We hope by this to lay a foundation for what elements of ML systems need to be included in System Cards, at which intervention points, and for which audiences. We also plan to tackle develop further transparency solutions, including data cards, method cards to account for different parts of the ML life cycle. Finally, we plan to explore how to integrate other aspects of and information about responsible AI elements of ML systems such as fairness, robustness, and accountability within our transparency solutions.

REFERENCES

Abubakar Abid, Ali Abdalla, Ali Abid, Dawood Khan, Abdulrahman Alfozan, and James Zou. 2019. Gradio: Hassle-free sharing and testing of ml models in the wild. *arXiv preprint arXiv:1906.02569* (2019).

Matthew Arnold, Rachel KE Bellamy, Michael Hind, Stephanie Houde, Sameep Mehta, Aleksandra Mojsilović, Ravi Nair, K Natesan Ramamurthy, Alexandra Olteanu, David Piorkowski, et al. 2019. FactSheets: Increasing trust in AI services through supplier's declarations of conformity. *IBM Journal of Research and Development* 63, 4/5 (2019), 6–1.

Emily M Bender and Batya Friedman. 2018. Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics* 6 (2018), 587–604.

Kasia S Chmielinski, Sarah Newman, Matt Taylor, Josh Joseph, Kemi Thomas, Jessica Yurkofsky, and Yue Chelsea Qiu. 2020. The dataset nutrition label (2nd Gen): Leveraging context to mitigate harms in artificial intelligence. In *NeurIPS Workshop on Dataset Curation and Security*.

Huanming Fang and Hui Miao. 2020. Introducing the Model Card Toolkit for Easier Model Transparency Reporting. Google AI Blog https://ai.googleblog.com/2020/07/introducing-model-card-toolkit-for.html. Accessed 2022-02-10.

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. 2021. Datasheets for Datasets. *Commun. ACM* 64, 12 (nov 2021), 86–92. https://doi.org/10.1145/3458723

Michael Hind, Stephanie Houde, Jacquelyn Martino, Aleksandra Mojsilovic, David Piorkowski, John Richards, and Kush R Varshney. 2020. Experiences with improving the transparency of AI models and services. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*. 1–8.

Sarah Holland, Ahmed Hosny, Sarah Newman, Joshua Joseph, and Kasia Chmielinski. 2018. The dataset nutrition label: A framework to drive higher data quality standards. *arXiv preprint arXiv:1805.03677* (2018).

Kenneth Holstein, Jennifer Wortman Vaughan, Hal Daumé III, Miro Dudik, and Hanna Wallach. 2019. Improving fairness in machine learning systems: What do industry practitioners need?. In *Proceedings of the 2019 CHI conference on human factors in computing systems*. 1–16.

P. M. Krafft, Meg Young, Michael Katell, Karen Huan, and Ghislain Bugingo. 2020. Defining AI in Policy versus Practice. In *Proceedings of the 2020 AAAI/ACM Conference on AI, Ethics, and Society (AIES '20*. 72–78.

Zachary C Lipton. 2018. The Mythos of Model Interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue* 16, 3 (2018), 31–57.

Nenad Medvidovic, David S Rosenblum, David F Redmiles, and Jason E Robbins. 2002. Modeling software architectures in the unified modeling language. *ACM Transactions on Software Engineering and Methodology (TOSEM)* 11, 1 (2002), 2–57.

Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. 2019. Model cards for model reporting. In *Proceedings of the conference on fairness, accountability, and transparency*. 220–229.

Oxford University Press. 2022. Lexico. https://www.lexico.com/en/definition/machine_learning.

Inioluwa Deborah Raji and Jingying Yang. 2019. ABOUT ML: Annotation and benchmarking on understanding and transparency of machine learning lifecycles. *arXiv preprint arXiv:1912.06166* (2019).

Adrian Weller. 2019. Transparency: motivations and challenges. In *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*. Springer, 23–40.

Kanit Wongsuphasawat, Daniel Smilkov, James Wexler, Jimbo Wilson, Dandelion Mane, Doug Fritz, Dilip Krishnan, Fernanda B Viégas, and Martin Wattenberg. 2017. Visualizing dataflow graphs of deep learning models in tensorflow. *IEEE transactions on visualization and computer graphics* 24, 1 (2017), 1–12.