A Multistakeholder Approach Towards Evaluating AI Transparency Mechanisms

Ana Lucic
Madhulika Srikumar
Partnership on AI
San Francisco, USA
ana@partnershiponai.org
madhu@partnershiponai.org

Ankur Taly
Google
San Francisco, USA
ataly@google.com

Umang Bhatt
University of Cambridge
Mozilla Foundation
Cambridge, UK
usb20@cam.ac.uk

Q. Vera Liao
IBM Research AI
New York City, USA
vera.liao@ibm.com

Alice Xiang
Sony AI
Tokyo, Japan
alice.xiang@sony.com

Maarten de Rijke
University of Amsterdam
Ahold Delhaize Research
Amsterdam, Netherlands
m.derijke@uva.nl

CHI'21, May 8-13, 2021, Online Virtual Conference
ACM 978-1-4503-6819-3/20/04.
https://doi.org/10.1145/3334480.XXXXXXX

Abstract
Given that there are a variety of stakeholders involved in, and affected by, decisions from machine learning (ML) models, it is important to consider that different stakeholders have different transparency needs [14]. Previous work found that the majority of deployed transparency mechanisms primarily serve technical stakeholders [2]. In our work, we want to investigate how well transparency mechanisms might work in practice for a more diverse set of stakeholders by conducting a large-scale, mixed-methods user study across a range of organizations, within a particular industry such as health care, criminal justice, or content moderation. In this paper, we outline the setup for our study.

Author Keywords
Transparency; Explainability; Interpretability

Introduction
There is an increased demand for transparency in artificial intelligence (AI) systems. So far, the explainable AI (XAI) community has primarily contributed computational methods for understanding predictions of machine learning (ML) models [6]. Such methods help users understand the rationale behind a model's behavior and typically range from global techniques that explain the entire model to local techniques that explain predictions from individual instances [6].

Previous work has shown that there exists a significant gap between research and deployment of transparency mechanisms for ML models: although many types of stakeholders are involved in the deployment of ML models, scenarios where transparency mechanisms are currently deployed are almost exclusively for stakeholders who build, validate, or deploy ML models [2]. This may be because most transparency mechanisms come in the form of local explanations, and although this is where the XAI community has focused a large portion of its efforts [6], perhaps it is not necessarily what stakeholders need in practice [1], other than ML engineers.

We argue that in order to meet the needs of non-technical stakeholders, we should think more broadly in terms of what transparency mechanisms are offered, going beyond explaining a model's *behavior* under certain conditions (i.e., its predictions), by also offering the *process* that went into building the model as an explanation. For example, a Model Card [13] might be more useful for an executive who is making a decision about whether or not to deploy a model at scale across their organization compared to seeing SHAP values [12] for five particular input instances.

To examine the effectiveness of various transparency mechanisms, we are building off of the previous work [2] conducted at our organization, the Partnership on AI: a global multistakeholder non-profit organization that aims to develop and share best practices for responsible use of AI.[1] This work was focused on taking stock of how transparency mechanisms are currently deployed across a range of industries and organizations. In our current work, we want to investigate how well these mechanisms work in practice by conducting a large-scale, mixed-methods, multistakeholder study on (a) providing meaningful explanations relevant to the specific needs of diverse stakeholders in different use cases, and (b) determining how these explanations should be evaluated. We want to focus on a particular industry (i.e., content moderation, criminal justice, health care, etc) to enable application-grounded evaluation [4], which involves real users conducting real tasks. We plan to examine the following research questions:

RQ1: What types of transparency mechanisms are most appropriate for different stakeholders in different use cases?

RQ2: How can we evaluate different types of transparency mechanisms in (a) objective terms such as a user's ability to perform a task using an explanation, and (b) subjective terms such as the impact on a user's trust in an AI system?

Preliminaries

A fundamental problem in the existing XAI literature is that the notions of transparency, interpretability and explainability are not well-defined and are often used interchangeably [8, 10]. We therefore explicitly define the terms used in our work as follows:

Transparency: providing insight into an ML model. Two possible forms of transparency include: *behavior-based transparency* and *process-based transparency*.

Behavior-based transparency: providing insight into how an ML model makes decisions, in a global or local manner, from an algorithmic or mathematical perspective. This is also sometimes referred to as *interpretability*. Some models are considered to be "inherently interpretable" (i.e., shallow decision trees or linear models with a small number of features), while others require *post-hoc* methods to generate these interpretations (i.e., SHAP values [12], LIME feature importances [15], counterfactual examples [11, 17]).

[1] https://www.partnershiponai.org

Process-based transparency: providing insight into the whole ML *modeling pipeline*, from development to production (i.e., models' intended use, data provenance, data collection, data splits for training and evaluation, team responsible for development and monitoring, evaluation metrics, reporting and visualization, etc. [5, 13]).

Explainability: translating transparency insights into something that is understandable to a human. Explainability may require either *behavior-based* or *process-based* transparency (or both), along with other information. It is conditioned on (a) the stakeholder's needs and characteristics, and (b) the use case in which it is deployed. For models that are considered to be "inherently interpretable" [16], this translation is still necessary since we need to decide how to present the information to the stakeholder.

Stakeholder: individuals who have a vested interest in the transparency of a system [2].

Use case: a particular context in which transparency is used or required.

Disentangling transparency and explainability in this way allows us to (a) present *process-based* transparency information as a potential explanation, and (b) separate the algorithmic component of generating model insights (i.e., the *interpretation*), from the form in which the information is presented to the user (i.e., the *explanation*). This allows us to have different explanations for different stakeholders, while using the same underlying information for transparency.

For example, in the context of *behavior-based* transparency, an end user might only be interested in a ranking of the most important features (e.g., ordered set of features based on mean SHAP [12] values), while an ML engineer might need more granular information (e.g., plots with individual SHAP values [12] for every sample, where each feature is shown on a separate plot). Although the underlying *interpretability* mechanism is the same (i.e., SHAP [12]), the resulting *explanations* are different, and can therefore be tailored to the stakeholders' needs.

Stakeholder-Informed Study Design

To construct our study, we plan to solicit input from relevant stakeholders in order to ensure the study represents tasks and subjective questions that reflect their values and transparency needs. This could take various forms including review panels, group workshops, and/or individual interviews with stakeholders, in particular those who interact with transparency mechanisms, in order to: (a) uncover common themes in participants' encounters and experiences with transparency techniques, and identify which type of explanations would be best suited for their use case, and (b) construct scenario-style sessions modeling real use cases to examine participants' transparency needs [3]

We will first design a pilot study, where stakeholders interact with various types of existing transparency mechanisms, both *behavior-* and *process-based*, and provide feedback on their experiences as they do so. The goal here would be to (a) identify which type of transparency mechanism is best suited for each stakeholder's particular use case, and (b) determine how best to translate the information provided by the underlying transparency mechanism into an explanation. These explanations would then be used as input for the user study outlined in the following section.

User Study

Based on the input we receive from stakeholders, we will design an application-grounded evaluation [4] study (i.e., a study with real users performing real tasks), in order to an-

swer **RQ1**. This would involve having stakeholders perform a set of industry-specific objective tasks, as well as answer some subjective questions about their experiences (e.g., Likert-scale). Examples of such tasks include forward or counterfactual simulations [4]. To elicit mental models [7] of how the ML model works, we could encourage stakeholders to think aloud while performing the tasks.

In order to test the effect on stakeholders' abilities to perform tasks, we would conduct a between-subject study, where all stakeholders are asked to perform the same tasks, but half would get some form of explanation while the other half would not. This would answer **RQ2a**.

To answer **RQ2b**, we would include a within-subject study for the stakeholders who had explanations by asking the same set of subjective questions (i.e., the Trust Scale in [7]) before and after completing the task, to see the effect that interacting with the explanations had on stakeholders' trust in, and satisfaction with, (a) the underlying model, and (b) the explanation.

Use Cases for Transparency

In this section, we outline some examples of use cases we hope to elicit by interacting with stakeholders within a particular industry. We plan to use the HCXAI workshop to narrow our focus regarding the possible use cases, based on prior work such as [9].

Interpreting individual predictions: providing an understanding of the salient factors for a particular prediction made by an ML model.

Gaining knowledge: generating new insights about the domain such as important decision factors or mechanisms [9], as well as understanding properties of the underlying dataset and task.

Aiding decisions: offering supporting evidence for a prediction, which allows the decision-maker to choose how to incorporate this information with their own knowledge in order to make a decision.

Suggesting interventions: suggesting appropriate interventions to the stakeholder in order to obtain a more favorable outcome, either from the model or in the real world.

Adapting system usage or control: allowing stakeholders to find the optimal ways to use the system, for example by adjusting their profiles or control settings.

Model improvement: offering insights that enable stakeholders to improve the model.

Model auditing: allowing investigation of concerns around model safety, ethics, and privacy.

Conclusion

Our work would be an important step in developing transparency mechanisms that are actually useful in practice to a diverse set of stakeholders. Model transparency is a multi-faceted problem, which does not have a single solution, and therefore the proposed solutions must be specific to both the use case and stakeholder involved. So far, we have had initial scoping conversations with over 15 partner organizations to gauge interest in the project, and are in the process of identifying an industry to center the study on. We are also designing a set of questions for soliciting input from stakeholders. We have also submitted a panel proposal to RightsCon[2] with the aim of facilitating a conversation between the XAI, HCI and human rights communities.

[2] https://www.rightscon.org/

Acknowledgments

This research was (partially) funded by the Hybrid Intelligence Center, a 10-year programme funded by the Dutch Ministry of Education, Culture and Science through the Netherlands Organisation for Scientific Research, https://hybrid-intelligence-centre.nl. All content represents the opinion of the authors, which is not necessarily shared or endorsed by their respective employers and/or sponsors.

REFERENCES

[1] Umang Bhatt, McKane Andrus, Adrian Weller, and Alice Xiang. 2020a. Machine learning explainability for external stakeholders. *arXiv preprint arXiv:2007.05408* (2020).

[2] Umang Bhatt, Alice Xiang, Shubham Sharma, Adrian Weller, Ankur Taly, Yunhan Jia, Joydeep Ghosh, Ruchir Puri, José M. F. Moura, and Peter Eckersley. 2020b. Explainable Machine Learning in Deployment. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. ACM, 648–657.

[3] Anna Brown, Alexandra Chouldechova, Emily Putnam-Hornstein, Andrew Tobin, and Rhema Vaithianathan. 2019. Toward Algorithmic Accountability in Public Services: A Qualitative Study of Affected Community Perspectives on Algorithmic Decision-Making in Child Welfare Services. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19)*. ACM, 1–12.

[4] Finale Doshi-Velez and Been Kim. 2017. Towards a Rigorous Science of Interpretable Machine Learning. *arXiv preprint arXiv:1702.08608v2* (2017).

[5] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. 2020. Datasheets for Datasets. *arXiv preprint arXiv:1803.09010* (2020).

[6] Riccardo Guidotti, Anna Monreale, Franco Turini, Dino Pedreschi, and Fosca Giannotti. 2018. A Survey of Methods for Explaining Black Box Models. *arXiv preprint arXiv:1802.01933* (2018).

[7] Robert R. Hoffman, Shane T. Mueller, Gary Klein, and Jordan Litman. 2018. Metrics for Explainable AI: Challenges and Prospects. *arXiv preprint arXiv:1812.04608* (2018).

[8] Harmanpreet Kaur, Harsha Nori, Samuel Jenkins, Rich Caruana, Hanna Wallach, and Jennifer Wortman Vaughan. 2020. Interpreting Interpretability: Understanding Data Scientists' Use of Interpretability Tools for Machine Learning. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. ACM, 1–14.

[9] Q Vera Liao, Daniel Gruen, and Sarah Miller. 2020. Questioning the AI: Informing Design Practices for Explainable AI User Experiences. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. ACM, 1–15.

[10] Zachary C. Lipton. 2016. The Mythos of Model Interpretability. *arXiv preprint arXiv:1606.03490* (June 2016).

[11] Ana Lucic, Harrie Oosterhuis, Hinda Haned, and Maarten de Rijke. 2019. FOCUS: Flexible Optimizable Counterfactual Explanations for Tree Ensembles. *arXiv preprint arXiv:1911.12199* (2019).

[12] Scott M Lundberg and Su-In Lee. 2017. A Unified Approach to Interpreting Model Predictions. In *NIPS*. Curran Associates, Inc., 4765–4774.

[13] 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 - FAT* '19*. ACM, 220–229.

[14] Sina Mohseni, Niloofar Zarei, and Eric D. Ragan. 2020. A Multidisciplinary Survey and Framework for Design and Evaluation of Explainable AI Systems. *arXiv preprint arXiv:1811.11839* (2020).

[15] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. Why Should I Trust You?: Explaining the Predictions of Any Classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 1135–1144.

[16] Cynthia Rudin. 2019. Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead. *Nature Machine Intelligence* 1 (2019), 206–215.

[17] Sandra Wachter, Brent Mittelstadt, and Chris Russell. 2018. Counterfactual Explanations without Opening the Black Box: Automated Decisions and the GDPR. *Harvard Journal of Law & Technology* 31, 2 (2018), 841–888.