Microsoft v2 Responsible AI Standard Reference Guide

In June 2022, we made our v2 Responsible AI Standard publicly available as part of our commitment to transparency, sharing our progress on our responsible AI journey, and raising awareness of our policies, programs, practices, and tools. We hope our approach and resources will be of value to other organizations. The Responsible AI Standard has had numerous expert contributions. The current version, version 2, was authored by a core working group of more than 30 individuals (research scientists, lawyers, designers, engineers, and other subject matter experts). In addition to the v2 Standard, the Office of Responsible AI, with support from the core working group, developed the supporting Impact Assessment and Impact Assessment Guide. Each of our core working group members brought their own disciplinary focus to this work and drew upon their experiences and existing responsible AI resources in their work on the Standard.

A key aim in making our Standard and accompanying resources available was to provide a full sharing of our learnings in the course of their development. Sharing a list of the academic resources that we have drawn inspiration from is an important part of our sharing, so we have compiled this reference guide, which cites academic resources that were influential in the construction of version 2 of the Responsible AI Standard. We have drawn on academic resources external to Microsoft, research published by Microsoft Research, and Microsoft Research done in collaboration with others. To more clearly document the transfer of knowledge from academia to industry, we have identified resources which are external to Microsoft with a green line (▌)to the left of the citation.

This guide is by no means exhaustive. In addition to the resources that we cite below, we have benefited immensely from our prior efforts and programs on responsible AI, including ideas reflected in an earlier version of the Standard. Our Responsible AI Standard also builds on Microsoft's decades of experience implementing technical standards (such as for privacy, security, and accessibility) and learnings from our own internal studies of how product teams understood and integrated proposed requirements for responsible AI into their workflows. We also have kept aware of regulatory proposals and their evolution while revising our Standard to create version 2, most notably with the release of the proposed EU AI Act. By making this list of resources available, we wish to acknowledge the contributions of others and hope that the efforts captured in the resources will be valuable to enabling responsible innovation in other organizations.

References

Saleema Amershi, Andrew Begel, Christian Bird, Robert DeLine, Harald Gall, Ece Kamar, Nachi Nagappan, Besmira Nushi, and Tom Zimmermann. 2019. Software Engineering for Machine Learning: A Case Study. In *2019 IEEE/ACM 41st International Conference on Software Engineering: Software Engineering in Practice (ICSE-SEIP)*, pp. 291-300.

> The large-scale study conducted with ML practitioners at Microsoft highlighted challenges and opportunities in several Responsible AI aspects throughout the whole ML lifecycle. Many of these challenges informed goal definitions across all RAI dimensions in the Standard.

Saleema Amershi, Dan Weld, Mihaela Vorvoreanu, Adam Fourney, Besmira Nushi, Penny Collisson, Jina Suh, Shamsi Iqbal, Paul N. Bennett, Kori Inkpen, Jaime Teevan, Ruth Kikin-Gil, and Eric Horvitz. 2019. Guidelines for Human-AI Interaction. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (CHI '19), Association for Computing Machinery, New York, NY, USA, 1–13. DOI:https://doi.org/10.1145/3290605.3300233

> This paper and the thinking behind the guidelines helped make clear how principles like Transparency and Reliability & Safety can be implemented in the UI, not only in documentation. It also made clear the importance of planning for failures.

Solon Barocas, Anhong Guo, Ece Kamar, Jacquelyn Krones, Meredith Ringel Morris, Jennifer Wortman Vaughan, W. Duncan Wadsworth, and Hanna Wallach. 2021. Designing Disaggregated Evaluations of AI Systems: Choices, Considerations, and Tradeoffs. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, ACM, Virtual Event USA, 368–378. DOI:https://doi.org/10.1145/3461702.3462610

> The requirements for two of our Fairness Goals (F1 and F2) outline how to perform a disaggregated evaluation to assess the fairness of AI system performance. This paper describes the choices, considerations, and tradeoffs of designing disaggregated evaluations and their impacts.

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. DOI:https://doi.org/10.1162/tacl_a_00041

> Goal A4: Data governance and management requires the documentation of training and testing data sets for AI systems. This value sensitive design paper describes one way to document data sets used to train natural language processing systems.

The Data Nutrition Project. https://datanutrition.org

> Goal A4: Data governance and management requires the documentation of training and testing data sets for AI systems. The Data Nutrition Project is building toward a Dataset Nutrition Label which displays standard quality measures for data sets used to train and test AI systems.

Thomas G. Dietterich and Eric J. Horvitz. 2015. Rise of concerns about AI: reflections and directions. *Communications of the ACM* 58, 10 (October 2015), 38–40. DOI:https://doi.org/10.1145/2770869

> This piece broadly informed and contributed to our approach to responsible AI at Microsoft.

Finale Doshi-Velez and Been Kim. 2017. Towards A Rigorous Science of Interpretable Machine Learning. DOI:https://doi.org/10.48550/arXiv.1702.08608

> In this position paper, which informed our thinking on intelligibility broadly, the author argues that despite increased interest in the area, there is very little consensus for what researchers mean by intelligibility, and outlines a direction for a rigorous approach to interpretability.

Batya Friedman and David G. Hendry. 2019. *Value Sensitive Design: Shaping Technology with Moral Imagination*. The MIT Press, Cambridge, MA.

> Value Sensitive Design provides foundational theory, method, and practice for surfacing human values in the technical design process. Value Sensitive Design's (VSD) stakeholder analysis provided a structure that we use in Microsoft's Responsible AI Impact Assessment to understand how an AI system may impact people, organizations, and society. VSD's framework of direct and indirect stakeholders also informed the stakeholder prompts we include in the Impact Assessment Guide. See also https://vsdesign.org/.

Timnit Gebru, Kate Crawford, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, and Hal Daumé III. 2021. Datasheets for Datasets. *Communications of the ACM,* Volume 64, Number 12, pages 86-92. https://cacm.acm.org/magazines/2021/12/256932-datasheets-for-datasets/fulltext

> Goal A4: Data governance and management requires the documentation of training and testing data sets for AI systems. Datasheets for Datasets is a general purpose template for documenting data sets used to train and test AI systems.

Eric Horvitz. 2014. One-Hundred Year Study on Artificial Intelligence: Reflections and Framing. Stanford University. https://ai100.stanford.edu/reflections-and-framing

This piece broadly informed and contributed to our approach to responsible AI at Microsoft.

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, Honolulu HI USA, 1–14. DOI:https://doi.org/10.1145/3313831.3376219

Goal T1: System intelligibility for decision making requires that we design AI systems to support stakeholders' decision-making needs and that we test that stakeholders can effectively interpret system responses. This paper influenced our decision to require evaluations in this Goal as it showed that our intuitions about what is intelligible to data scientists can be wrong.

Zachary C. Lipton. 2018. The mythos of model interpretability. *Communications of the ACM* 61, 10 (September 2018), 36–43. DOI:https://doi.org/10.1145/3233231

In this position paper, which influenced our thinking on intelligibility broadly, the author argues that model interpretability is an important yet slippery concept, and analyzes the literature to date to demonstrate the conflicting ways in which intelligibility researchers conceptualize and advocate for model interpretability.

Michael A. Madaio, Luke Stark, Jennifer Wortman Vaughan, and Hanna Wallach. 2020. Co-Designing Checklists to Understand Organizational Challenges and Opportunities around Fairness in AI. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, ACM, Honolulu HI USA, 1–14. DOI:https://doi.org/10.1145/3313831.3376445

In this work the authors identify challenges and opportunities for fairness in AI systems and co-design a fairness checklist with AI practitioners. Many of the elements in the fairness checklist that was co-designed through this work became requirements of the Responsible AI Standard.

Ram Shankar Siva Kumar, Jeffrey Snover, David O'Brien, Kendra Albert, and Salome Viljoen. 2019. Failure Modes in Machine Learning. Failure Modes in Machine Learning - Security documentation | Microsoft Learn

This guide summarizes recent work on adversarial and non-adversarial failure modes for machine learning systems. Considering both adversarial and non-adversarial failures is a core part of our approach to upholding Reliability, Safety and Security.

Tim Miller. 2018. Explanation in Artificial Intelligence: Insights from the Social Sciences. DOI:https://doi.org/10.48550/arXiv.1706.07269

In this position paper, which influenced our thinking on interpretability broadly, the author argues that the field of explainable artificial intelligence has to date relied on intuitive notions of what makes a 'good' explanation. Instead, the field should build on existing research on explanations in fields such as philosophy, cognitive psychology, cognitive science, and social psychology.

Cecily Morrison, Edward Cutrell, Martin Grayson, Anja Thieme, Alex Taylor, Geert Roumen, Camilla Longden, Sebastian Tschiatschek, Rita Faia Marques, and Abigail Sellen. 2021. Social Sensemaking with AI: Designing an Open-ended AI experience with a Blind Child. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pp. 1-14. 2021.

In this work the authors co-design an AI system with a blind child and reflect on working with the capabilities of an AI system. This work strengthened our perspective that the capabilities of a system arise from the human-AI interaction and not the system alone.

Don A. Norman. 1987. Some observations on mental models. In W.A.S. Buxton & R.M. Baecker (Eds.), *Human-computer interaction: a multidisciplinary approach* (pp. 241-244). Morgan Kaufmann Publishers Inc.

> This work emphasizes the need for system design to support a workable mental model for users, of particular focus for both Goal T1: System intelligibility for decision making and A5 Human oversight and control.

Besmira Nushi, Ece Kamar, and Eric Horvitz. 2018. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing* (Vol. 6, pp. 126-135).

> The approach described in the paper was the foundation upon which the Error Analysis tool was designed and conceptualized. The tool enables ML practitioners to conduct disaggregated evaluation and identify failure modes. Such activities have a direct impact on supporting elements (Tools and Practices) discussed in the RAI Standard related to Fairness, Reliability, and Safety.

Forough Poursabzi-Sangdeh, Daniel G Goldstein, Jake M Hofman, Jennifer Wortman Vaughan, and Hanna Wallach. 2021. Manipulating and Measuring Model Interpretability. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, ACM, Yokohama Japan, 1–52. DOI:https://doi.org/10.1145/3411764.3445315

> Goal T1: System intelligibility for decision making requires that we design AI systems to support stakeholders' decision-making needs and that we test that stakeholders can effectively interpret system responses. This paper influenced our decision to require evaluations in this Goal as it showed that our intuitions about what is intelligible to non-expert users can be wrong.

Bogdana Rakova, Jingying Yang, Henriette Cramer, and Rumman Chowdhury. 2021. Where Responsible AI meets Reality: Practitioner Perspectives on Enablers for Shifting Organizational Practices. Proc. *ACM Hum.-Comput. Interact. 5, CSCW1,* Article 7 (April 2021), 23 pages. DOI:https://doi.org/10.1145/3449081

> This paper articulates practical challenges companies are facing in implementing responsible AI and emphasizes the importance of organizational structures to support and streamline responsible AI work throughout product development.

Jennifer Wortman Vaughan and Hanna Wallach. 2021. A human-centered agenda for intelligible machine learning. In *In Machines We Trust: Perspectives on Dependable AI,* edited by Marcello Pelillo and Teresa Scantamburlo. MIT Press. http://www.jennwv.com/papers/intel-chapter.pdf

> This book chapter establishes a stance on intelligibility, different stakeholders and goals, and the importance of human-centered evaluation. It reflects our broader thinking in this space which made its way into the Standard.

Sebastian Hallensleben, Carla Hustedt, Lajla Fetic Torsten Fleischer, Paul Grünke, Thilo Hagendorff, Marc Hauer, Andreas Hauschke, Jessica Heesen, Michael Herrmann, Rafaela Hillerbrand, Christoph Hubig, Andreas Kaminski, Tobias Krafft, Wulf Loh, Philipp Otto, and Michael Puntschuh. 2020. From Principles to Practice: An interdisciplinary framework to operationalise AI ethics. Report of the AIEI Group. VDE Association for Electrical, Electronic & Information Technologies and Bertelsmann Stiftung. https://www.ai-ethics-impact.org/en

> The multi-tiered framework proposed by the "VCIO model" in this report influenced our design of the structure of the RAI Standard, with its Principles, Goals, Requirements, and Tools & Practices.