NIST Workshop on AI Measurement and Evaluation Agenda

All times EDT (UTC-4)

Day 1: Tuesday June 15, 2021

Time Start	Time End	Topic	Speaker(s)
11:00 AM	11:20 AM	Welcome, Workshop Goals & Logistics Overview	**Elham Tabassi** (Chief of Staff, Information Technology Laboratory, NIST)
11:20 AM	11:50 AM	A National Security Perspective on AI Measurement and Evaluation	**Jason Matheny** (Deputy Assistant to the President for Technology and National Security; Deputy Director for National Security in the White House Office of Science and Technology Policy; and Coordinator for Technology and National Security at the National Security Council)
11:50 AM	12:00 PM	Break	
12:00 PM	1:30 PM	Panel 1: Measuring with Purpose Discussion of the needs for and uses of AI evaluation outputs and their role in driving down-stream processes, including the requirements and properties important for an AI evaluation to possess in order to be fit for the intended uses. Identification of areas for which current measurement and evaluation approaches are insufficient or do not exist, where further AI metrology research would be beneficial.	**Moderator:** Tess DeBlanc-Knowles (White House Office of Science and Technology Policy) **Panelists:** Jack Clark (Anthropic) Michael Hind (IBM Research) Chuck Howell (MITRE) Jane Pinelis (Test and Evaluation of AI/ML at DoD Joint Artificial Intelligence Center) Salvatore Scalzo (European Commission) Bill Scherlis (DARPA)
1:30 PM	1:45 PM	Break and Discussion Time	
1:45 PM	2:15 PM	Panel 2: Overview of Past & Current Evaluations Overview of the evaluation-driven research paradigm that has been used at NIST to evaluate AI systems, with a description of the various styles of evaluations, as well as examples of some of the AI	**Moderator:** Mark Przybocki (NIST) **Panelists:** Peter Bajcsy (NIST) Jonathan Fiscus (NIST) Jonathon Phillips (NIST) Michael Sharp (NIST) Ellen Voorhees (NIST) Megan Zimmerman (NIST)

		measurement and evaluation activities conducted at NIST.	
2:15 PM	2:45 PM	Panel 3: Discussion of NIST/Community Future Work Discussion of the limitations of current AI measurement and evaluation activities that prevent them from addressing all the needs for AI measurement and evaluation, and future plans for NIST to address these limitations together with the research community.	
2:45 PM	3:00 PM	Break	
3:00 PM	4:00 PM	Panel 4: Evaluating AI during Operation Discussion of AI evaluation in production/operational environments, including topics drawn from: MLOps; Operational evaluation metrics/Business metrics; Model quality/Data drift with online data; Latency, throughput, and scalability issues; Adversarial attacks and robustness to corruptions/perturbations; Governance and regulatory compliance.	**Moderator:** Antonio Moretti (Walmart) **Panelists:** Clarence Agbi (Brex) Sergey Karayev (Turnitin) Josh Tobin (Gantry)
4:00 PM	4:15 PM	Closing Remarks	NIST Workshop Organizing Committee
4:15 PM	5:00 PM	After Hours: Slack with NIST staff	

Day 2: Wednesday June 16, 2021

Time Start	**Time End**	**Topic**	
11:00 AM	11:30 AM	Keynote	**Fei-Fei Li** (Sequoia Professor, Stanford University; Co-Director of Stanford's Human-Centered AI Institute)
11:30 AM	12:30 PM	Panel 5: Evaluation Design Process	**Moderator:** Nicholas Carlini (Google Brain)

		Discussion of the processes and procedures for designing evaluations of AI systems, including: the high-level considerations and decisions that must be made in order to design and implement effective evaluations; the components of and relationships between the various evaluation design elements; and the role of the applications and overall evaluation goals in evaluation design.	**Panelists:** Matthias Hein (University of Tübingen) Deborah Raji (Mozilla Foundation) Shibani Santurkar (MIT / Stanford) Ludwig Schmidt (Toyota Research / UW)
12:30 PM	12:45 PM	Break	
12:45 PM	1:45 PM	Panel 6: Metrics and Measurement Methods Discussion of: the properties of an AI system that can/should be measured, and which properties have/lack metrics and measurement methods; the different measurement methods that are used to measure AI and their strengths/limitations; the different types and uses of metrics, and the various properties that a metric can poses; the impacts of the chosen metrics and measurements methods have on an evaluation; when is it important to have glass box access to AI systems for evaluation, and when the design/approach taken by an AI system influences the choice of metrics/measurement methods.	**Moderator:** Craig Greenberg (NIST) **Panelists:** José Hernández-Orallo (Universitat Politècnica de València) Douglas Reynolds (NSA / MIT Lincoln Laboratory) Sameer Singh (UCI)
1:45 PM	2:00 PM	Break	
2:00 PM	3:00 PM	Panel 7: Data and Data Sets Data collection methods and dataset design for AI system measurement and evaluation, along with discussions drawing from the following topics: approaches for data annotation/labeling; uncertain, missing, or non-existence of ground truth; how much data is	**Moderator:** Aleksander Mądry (MIT) **Panelists:** Marzyeh Ghassemi (U Toronto/MIT) Tom Goldstein (UMD) Emre Kiciman (MSR) Nicolas Papernot (U Toronto)

		necessary; needs for and uses of simulated/generated data; roles of common datasets in research; repurposing of data; ethical and privacy considerations; et al.	
3:00 PM	3:15 PM	Break	
3:15 PM	4:15 PM	Panel 8: Limitations, Challenges and Future Directions of Evaluation Discussion of the limitations, challenges, shortcomings, and future directions for the evaluation and measurement of AI, including the new or emerging evaluation paradigms, the ability/inability to generalize evaluation results and its policy implications. Needs and plans for improvements to existing measurement and evaluation activities as well as the creation of new AI evaluation challenge problems and measurement research.	**Moderator:** Soheil Feizi (UMD) **Panelists:** Kamalika Chaudhuri (UCSD) Eric Horvitz (MSR) Percy Liang (Stanford) Chris Meserole (Brookings) Daniela Rus (MIT)
4:15 PM	4:30 PM	Break and Slack Discussion Time	
4:30 PM	5:00 PM	Closing Remarks	NIST Workshop Organizing Committee
5:00 PM	5:30 PM	After Hours: Slack with NIST staff	

Day 3: Thursday June 17, 2021

Time Start	**Time End**	**Topic**	
11:00 AM	11:30 AM	AI test and evaluation from National AI Initiative Perspective	**Lynne Parker** (Director, National AI Initiative Office, White House Office of Science and Technology Policy)
11:30 AM	12:30 PM	Panel 9: Measuring Concepts that are Complex, Contextual and Abstract Discussion of the challenges and approaches for measuring AI system characteristics that are complex, contextual, and/or	**Moderator:** Ellen Voorhees (NIST) **Panelists:** Lora Aroyo (Google) Ben Carterette (Spotify) David Ferrucci (Elemental Cognition)

		abstract, or are otherwise difficult to quantify (such as explainability, bias, trustworthiness, safety, etc.) including the role that descriptive and/or qualitative measurements should play in these cases.	
12:30 PM	12:45 PM	Break	
12:45 PM	1:45 PM	Panel 10: Measuring with Humans in the Mix Discussion of the measurement and evaluation of AI systems that work in cooperation with humans, including the roles and relationships between the AI systems and the humans, and the challenges of and approaches to measurement and evaluation when humans and AI systems are involved.	**Moderator:** Margaret Burnett (OSU) **Panelists:** Rachel Bellamy (IBM) Madeleine Clare Elish (Google) Robert Hoffman (IHMC)
1:45 PM	2:00 PM	Break	
2:00 PM	3:00 PM	Panel 11: Software Infrastructure Overview, Existing Tools and Future Desires Discussion of the landscape, challenges, and needs of developing tools and infrastructure for the particular purpose of measuring, testing, and evaluating AI systems.	**Moderator:** Harold Booth (NIST) **Panelists:** Pin-Yu Chen (IBM) Harsha Nori (Microsoft) David Pitman (Google)
3:00 PM	3:15 PM	Break and Discussion Time	
3:15 PM	4:15 PM	Panel 12: Practical Considerations and Best Practices for Measurement and Evaluation Discussion of the practical considerations and concrete best practices for the measurement and evaluation of AI-based systems, including the testing and evaluation strategies that can be used to mitigate privacy loss or intellectual property exposure in AI testing.	**Moderator:** William Streilein (MIT Lincoln Laboratory) **Panelists:** Matt Gaston (SEI Emerging Technology Center, CMU) Sven Krasser (CrowdStrike) Sanjeev Mohindra (MIT Lincoln Laboratory) Jane Pinelis (Test and Evaluation of AI/ML at DoD Joint Artificial Intelligence Center) Richard Tatum (CIV USN NAVSURFWARCEN PNC FL)
4:15 PM	4:30 PM	Break	

4:30 PM	5:00 PM	NIST: Workshop Debrief and Next Steps	NIST Workshop Organizing Committee