Assessing and Improving AI Trustworthiness: Current Contexts, Potential Paths
Workshop Agenda

March 3: 12:00 - 5:15pm Eastern

12:00pm - 12:05pm - Welcome

12:05pm - 12:15pm - Introductory Remarks by Elham Tabassi, Chief of Staff, Information Technology Laboratory at the National Institute of Standards and Technology

1. **12:15pm - 1:45pm - AI Trustworthiness in Context**
   Discussing how system builders and assessors construe the various aspects of AI trustworthiness, including common performance metrics or assessment methods. This session will examine this topic for three high stakes areas where AI is being deployed -- financial services, healthcare, and transportation -- pairing panelists who work on assessment standards with those who actively build and assess production AI systems.

   Financial Services
   - Mr. David Palmer, Senior Supervisory Financial Analyst, Federal Reserve Board
   - Dr. Agus Sudjianto, Executive Vice President and Head of Corporate Model Risk, Wells Fargo

   Healthcare
   - Mr. Bakul Patel, Director of Digital Health, Food and Drug Administration
   - Prof. Mark Sendak, Population Health and Data Science Lead, Duke Institute for Health Innovation

   Transportation
   - Mr. Chris Hart, Founder, Hart Solutions LLC and former Chair, National Transportation Safety Board
   - Ms. Deborah Prince, Standards Program Manager, Underwriters Laboratories

   Committee Moderators
   - Prof. Anupam Datta, Carnegie Mellon University
   - Prof. Deirdre Mulligan, University of California, Berkeley

2. **1:45pm – 2:15pm – Keynote Address by Dr. Eric Horvitz, National Security Commission on Artificial Intelligence**

3. **2:15pm - 3:45pm - Attributes of AI Trustworthiness: Robust, Explainable, Generalizable**
   This session will go deeper into several component attributes that contribute to system trustworthiness. The panelists will discuss how system designers enhance

trustworthiness by incorporating explainability, robustness, and generalizability, and how they approach tradeoffs among these various system qualities, and the importance of being human-centered.

Speakers

- Prof. Katherine Heller, Duke University, and Research Scientist, Google Brain
- Prof. Aleksander Madry, Massachusetts Institute of Technology
- Dr. Martin Wattenberg, Co-lead, Google People and AI Research initiative (PAIR)

Committee Moderators

- Dr. Krishnamurthy (Dj) Dvijotham, DeepMind
- Prof. Ben Shneiderman, University of Maryland

4. **3:45pm - 5:15pm - Attributes of AI Trustworthiness: Fair, Private, Contestable**
   This session will go deeper into several component attributes that contribute to system trustworthiness. The panelists will discuss how privacy, fairness, and the ability to challenge the outputs of AI systems contribute to trustworthiness, and how they approach tradeoffs among these system qualities.

   Speakers

   - Prof. Tad Hirsch, Northeastern University
   - Prof. Michael Kearns, University of Pennsylvania
   - Dr. Ilya Mironov, Research Scientist, Facebook AI
   - Dr. Jenn Wortman Vaughan, Senior Principal Researcher, Microsoft

   Committee Moderators

   - Prof. Aleksander Madry, Massachusetts Institute of Technology
   - Prof. Deirdre Mulligan, University of California, Berkeley

March 4: 12:00pm - 4:00pm ET

5. **12:00pm - 1:30pm - Measurement for AI Systems**
   Examining the role of the measurement sciences community, including academia, independent observers, and NIST, in developing tools to improve various aspects of AI performance. This will explore both past NIST work, to illustrate the typical activities of that organization, and ongoing work by academic and independent researchers to create benchmarks, heuristics, and other tools to allow for better assessment and comparison of system performance for different aspects of trustworthiness.

   Speakers

   - Prof. Alexandra Chouldechova, Carnegie Mellon University
   - Mr. Nicolas Economou, Chief Executive Officer, H5 and Law Committee Chair, IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems

- Dr. Joaquin Quiñonero Candela, Director of Applied Machine Learning, Facebook
- Dr. Sriram Rajamani, Distinguished Scientist and Managing Director, Microsoft Research India
- Prof. Suchi Saria, Johns Hopkins University

Committee Moderators

- Dr. Susan Dumais, Microsoft Research
- Prof. Jeannette Wing, Columbia University

1:30pm – 2:00pm – Break

6. **2:00pm – 4:00pm - Workshop Synthesis and Outcomes**
   Identifying potential 'next steps' that partners in academia and industry can undertake to use their different areas of expertise to build upon existing tools and promote a shared toolkit for AI trustworthiness assessment and formulating recommendations for future work for NIST and other public agencies.

   Sponsor Representatives

   - Chuck Romine, NIST
   - Elham Tabassi, NIST

   Workshop Committee

   - Committee Chair: Prof. Anupam Datta, Carnegie Mellon University
   - Dr. Susan Dumais, Microsoft Research
   - Dr. Dj Dvijotham, DeepMind
   - Prof. Aleksander Madry, Massachusetts Institute of Technology
   - Prof. Deirdre Mulligan, University of California, Berkeley
   - Prof. Ben Shneiderman, University of Maryland
   - Prof. Jeannette Wing, Columbia University