NATIONAL ACADEMIES Sciences Engineering Medicine

Division on Engineering and Physical Sciences

Human and Organizational Factors in AI Risk Management - A Workshop

[Virtual]

The goal for this meeting is to identify and explore approaches to addressing human and organizational risks in AI systems. The emphasis will be on approaches that could be included in a more detailed guidance document complementing the recently issued "AI Risk Management Framework" from the National Institute of Standards and Technology (NIST). Areas of interest in the context of the NIST framework include human insights about AI-produced output and human oversight of AI systems and their operation in real-world environments.

Join us online on **June 20 from 10:30am-3:45pm ET** to discuss how AI tool evaluations might be scoped, evaluation methods and mechanisms, and the governance of evaluations.

THURSDAY, JUNE 20, 2024

10:30-10:55 am ET

Welcome by Workshop Co-Chairs and Committee
Mona Sloane (co-chair), Assistant Professor, University of Virginia
Ben Shneiderman (co-chair), NAE, Professor Emeritus, University of Maryland
Abigail Jacobs, Assistant Professor of Information, Assistant Professor of Complex Systems, University of Michigan

10:55-11:15 am ET

Opening Remarks
Alexandra Givens, CEO, Center for Democracy & Technology

11:15-12:15 pm ET

I. Scoping Evaluations

Moderator: **William Isaac**, Principal Scientist and Head of Ethics Research, Google DeepMind

Speakers:

- **Rishi Bommasani,** Stanford University
- **Hanna Wallach,** Partner Research Manager, Microsoft Research
- **Laura Weidinger**, Senior Research Scientist, Google DeepMind
- **Miranda Bogen**, AI Governance Lab Director, Center for Democracy & Technology

12:15-12:45 pm ET

Lunch

12:45 am-2:00 pm ET

II. Evaluation Methods and Mechanisms

Moderator: **Hoda Heidari**, K&L Gates Career Development Assistant Professor in Ethics and Computational Technologies, Carnegie Mellon University

Speakers:

- **Lama Ahmad,** Policy Research, OpenAI
- **Kenneth Holstein**, Assistant Professor in the Human-Computer Interaction Institute, Carnegie Mellon University
- **Margaret Mitchell**, Researcher and Chief Ethics Scientist, Hugging Face
- **Arvind Narayanan**, Professor of Computer Science, Princeton University

2:00-2:30 pm ET **Break**

2:30-3:30 pm ET **III. Governance of Evaluation**

Moderator: **Solon Barocas**, Principal Researcher, Microsoft Research; Adjunct Assistant Professor of Information Science, Cornell

Speakers:

- **Daniel Ho,** William Benjamin Scott and Luna M. Scott Professor of Law, Professor of Political Science, Professor of Computer Science (by courtesy), Senior Fellow at the Stanford Institute for Human-Centered Artificial Intelligence, Senior Fellow at the Stanford Institute for Economic Policy Research, and Director of the Regulation, Evaluation, and Governance Lab, Stanford University
- **Jacob Metcalf,** Director of the AI on the Ground Initiative, Data & Society
- **Diane Staheli**, Assistant Director of AI Applications, White House Office of Science and Technology Policy

3:30-3:45 pm ET **Closing Remarks**

Abigail Jacobs, Assistant Professor of Information, Assistant Professor of Complex Systems, University of Michigan
Solon Barocas, Principal Researcher, Microsoft Research; Adjunct Assistant Professor of Information Science, Cornell

3:45 pm ET **Workshop Event 2 Adjourn**

Speaker Biographical Sketches

Lama Ahmad is a Policy Researcher at OpenAI, leading red teaming and researcher access efforts. Her work focuses on evaluating the socio-technical impact of AI systems on society. Prior to OpenAI, Lama was at Meta, assessing the impact of social media on elections and democracy.

Miranda Bogen is the founding director of the AI Governance Lab at the Center for Democracy & Technology. Building on CDT's decades of leadership fighting to advance civil rights and civil liberties in the digital age, the Lab provides public interest expertise in rapidly developing policy and technical conversations around artificial intelligence, advancing the interests of individuals whose lives and rights are impacted by AI. An AI policy expert and responsible AI practitioner, Miranda has led work at the intersection of policy and AI fairness and governance in senior roles in industry and civil society. She served as co-chair of the Fairness, Transparency, and Accountability Working Group at the Partnership on AI, conducted foundational research at the intersection of machine learning and civil rights at Upturn, and most recently guided strategy and implementation of responsible AI practices at Meta.

Rishi Bommasani is the Society Lead at the Stanford Center for Research on Foundation Models, where he researches the societal impact of foundation models and leads the Center's policy initiatives. His work has been covered in The Atlantic, Bloomberg, Nature, The New York Times, The Wall Street Journal, and The Washington Post. Rishi is completing his PhD at Stanford Computer Science, advised by Percy Liang and Dan Jurafsky and supported by the NSF Graduate Research Fellowship and Stanford Lieberman Fellowship.

Alexandra Reeve Givens is the CEO of the Center for Democracy & Technology, a nonpartisan, nonprofit organization fighting to protect civil rights and civil liberties in the digital age. She is a frequent public commentator on ways to protect users' online privacy and access to information, and to ensure emerging technologies advance human rights and democratic values. Alex previously served in the United States Senate, as the chief counsel on the Senate Judiciary Committee covering innovation and consumer protection, as well as as a litigator in private practice. She taught for nine years as an adjunct professor at Columbia Law and Georgetown Law. She holds a B.A. from Yale University and a J.D. from Columbia University School of Law.

Dr. Daniel E. Ho is the William Benjamin Scott and Luna M. Scott Professor of Law, Professor of Political Science, Professor of Computer Science (by courtesy), Senior Fellow at Stanford's Institute for Human-Centered Artificial Intelligence, and Senior Fellow at the Stanford Institute for Economic Policy Research at Stanford University. He is a Faculty Fellow at the Center for Advanced Study in the Behavioral Sciences and is Director of the Regulation, Evaluation, and Governance Lab (RegLab). Ho serves on the National Artificial Intelligence Advisory Committee (NAIAC), advising the White House on artificial intelligence, as Senior Advisor on Responsible AI at the U.S. Department of Labor, and as a Public Member of the Administrative Conference of the United States (ACUS). He is an elected member of the American Academy of Arts and Sciences. He received his J.D. from Yale Law School and Ph.D. from Harvard University and clerked for Judge Stephen F. Williams on the U.S. Court of Appeals, District of Columbia Circuit.

Dr. Ken Holstein is an Assistant Professor at Carnegie Mellon University's Human-Computer Interaction Institute, where he directs Co-{Augmentation, Learning, & AI} (CoALA) Lab. His research is broadly concerned with enabling more participatory, worker-centered, and community-driven approaches to AI design and evaluation. Towards this goal, researchers at the CoALA Lab develop new methods and tools to incorporate diverse human expertise across the AI development lifecycle. His group's research has received several Best Paper recognitions at top-tier venues in human-centered AI, and has been featured by outlets such as PBS, The Guardian, Wired, Forbes, and The Boston Globe.

Dr. Jacob (Jake) **Metcalf** is a researcher at Data & Society, where he leads the AI on the Ground Initiative, and works on an NSF-funded multisite project, Pervasive Data Ethics for Computational

Research (PERVADE). For this project, he studies how data ethics practices are emerging in environments that have not previously grappled with research ethics, such as industry, IRBs, and civil society organizations. His recent work has focused on the new organizational roles that have developed around AI ethics in tech companies. Jake's consulting firm, Ethical Resolve, provides a range of ethics services, helping clients to make well-informed, consistent, actionable, and timely business decisions that reflect their values. He also serves as the Ethics Subgroup Chair for the IEEE P7000 Standard.

Dr. Margaret Mitchell is a researcher focused on the ins and outs of machine learning and ethics-informed AI development in tech. She has published around 100 papers on natural language generation, assistive technology, computer vision, and AI ethics, and holds multiple patents in the areas of conversation generation and sentiment classification. She has recently received recognition as one of Time's Most Influential People of 2023. She currently works at Hugging Face as Chief Ethics Scientist, driving forward work in the ML development ecosystem, ML data governance, AI evaluation, and AI ethics. She previously worked at Google AI as a Staff Research Scientist, where she founded and co-led Google's Ethical AI group, focused on foundational AI ethics research and operationalizing AI ethics Google-internally. Before joining Google, she was a researcher at Microsoft Research, focused on computer vision-to-language generation; and was a postdoc at Johns Hopkins, focused on Bayesian modeling and information extraction. She holds a PhD in Computer Science from the University of Aberdeen and a Master's in computational linguistics from the University of Washington. While earning her degrees, she also worked from 2005-2012 on machine learning, neurological disorders, and assistive technology at Oregon Health and Science University. She has spearheaded a number of workshops and initiatives at the intersections of diversity, inclusion, computer science, and ethics. Her work has received awards from Secretary of Defense Ash Carter and the American Foundation for the Blind, and has been implemented by multiple technology companies.

Dr. Deirdre K. Mulligan serves as Principal Deputy U.S. Chief Technology Officer at the White House Office of Science and Technology Policy. In this role, Mulligan leads the Tech Division within OSTP, working to ensure that digital technologies benefit all Americans and advance democratic values. Mulligan is a Professor at the School of Information at UC Berkeley and a Faculty Director of the Berkeley Center for Law & Technology. Her research focuses on protecting values including privacy, equity, and freedom of expression in sociotechnical systems. Prior to joining the School of Information, Mulligan was the first Director of the Samuelson Law and Technology Clinic and a Clinical Professor at UC Berkeley School of Law. She is a founding board member of the Partnership on AI, a founding member of the Global Network Initiative, and a former Commissioner on the Oakland Privacy Advisory Commission. She helped start the Center for Democracy and Technology, where she worked on key tech policy issues during the emergence of the commercial internet.

Dr. Arvind Narayanan is a professor of computer science at Princeton and the director of the Center for Information Technology Policy. He co-authored a textbook on fairness and machine learning and is currently co-authoring a book on AI snake oil. He led the Princeton Web Transparency and Accountability Project to uncover how companies collect and use our personal information. His work was among the first to show how machine learning reflects cultural stereotypes, and his doctoral research showed the fundamental limits of de-identification. Narayanan is a recipient of the Presidential Early Career Award for Scientists and Engineers (PECASE).

Dr. Hanna Wallach is a partner research manager at Microsoft Research New York City. She is also an adjunct professor in the College of Information and Computer Sciences at UMass Amherst and a member of UMass Amherst's Computational Social Science Institute. Her research focuses on issues of fairness, accountability, transparency, and ethics as they relate to AI and machine learning. She collaborates with researchers from machine learning, natural language processing, human–computer interaction, and science and technology studies, as well as lawyers and policy makers; her research integrates both qualitative and quantitative perspectives. Previously, she developed machine learning and natural language processing methods for analyzing the structure, content, and dynamics of social processes. She currently serves on the NeurIPS Executive Board, the ICML Board, the FAccT Steering Committee, the WiML Senior Advisory Council, and the WiNLP Advisory Board. Hanna is committed to increasing diversity in computing and has worked for almost two decades to address the underrepresentation of women, in particular. To that end, she co-founded two projects—the first of their kind—to increase women's involvement in free and open-source software development: Debian Women and the GNOME Women's Summer Outreach Program (now Outreachy). She also co-founded the WiML Workshop. She holds a BA in computer science from the University of Cambridge, an MS in cognitive science from the University of Edinburgh, and a PhD in machine learning from the University of Cambridge.

Dr. Laura Weidinger is a Staff Research Scientist at Google DeepMind, where she leads research on ethics and safety evaluation of AI systems. Laura's work has focused on taxonomising, measuring and mitigating risks from generative AI systems and on developing novel evaluation approaches for AI. Previously, Laura worked in cognitive science research and as policy advisor at UK and EU levels. She holds degrees from Humboldt Universität Berlin and University of Cambridge.