TRAILS Con 2026: Evaluating AI

Conference Agenda

March 4, 2026	
9 – 9:30 a.m.	**Welcome Remarks** ***Speakers:*** *Hal Daumé III (TRAILS Institute Director, University of Maryland); Zoe Szajnfarber (Senior Advisor to the President on AI Strategy, George Washington University)*
9:30 - 10:00 a.m.	**Fireside Chat - Michael Madaio (Google)**
10 - 11 a.m.	**Engaging Participation in AI Evaluation (Panel)** Evaluating AI systems and outputs takes a broad range of expertise. How can we best engage broad and diverse publics in AI evaluation? This panel will gather experts exploring exciting new methods for public engagement in evaluation. ***Panelists****: Razvan Amironesei (NIST); Michael Ekstrand (Drexel University); Aruneesh Salhotra (OWASP Foundation)* ***Moderator****: Katie Shilton (University of Maryland)*
11 a.m. – 12:00 p.m.	**Measuring Impact of AI Use on Productivity (Panel)** ***Panelists:*** *Ray Bell (State of Maryland), Todd Marks (Mindgrub*

	Technologies) ***Moderator**: Tom Goldstein (University of Maryland)*
12:15 – 1:15 p.m.	**Networking Lunch and GW TAI - TRAILS Hackathon Showcase**
1:15 – 2:25 p.m.	**Sensemaking: The Link Between Evaluation and Trust (Panel)** ***Panelists:** Bruce Gellin (Georgetown University Global Health Institute); Marine Carpuat (University of Maryland); Reva Schwartz (Civitaas)* ***Moderator**: David Broniatowski (GW University)*
2:30 – 3:40 p.m.	**An Assessment of Voluntary AI Governance at Various Levels of Government (Panel)** International organizations, nations, states and firms have developed principles, standards, rules, guidelines, laws, and accountability mechanisms such as evaluations to govern AI. Given what we know about known risks, how should we govern AI? Should it be through transparency, explainability, regulation, and governance data? Is AI ungovernable or are we simply in the early phases of governing AI? What roles should interoperability (by definition global) and competition policy (mainly national) play? Who should decide on how to evaluate and what are acceptable evaluations? ***Panelists:** Ben Brake (DOT Europe); Jesse Dunietz (National Institute of Standards and Technology); Kevin Klyman (Google); Randi Michel (CA Advisor for AI, Office of Governor Gavin Newsom); Brandie Nonnecke (Americans for Responsible Innovation);* ***Moderator**: Susan Ariel Aaronson (GW University)*
3:40 – 4 p.m.	**Networking Break**
4 – 4:30 p.m.	**Fireside Chat**

4:30 – 5:15 p.m.	**Defining the Future of AI Evaluation (Panel)** **Senior TRAILS researchers come together to reflect on where** AI evaluation is headed and what it will take to assess AI systems in ways that actually matter in real-world use. The conversation will span technical, social, and governance perspectives, highlighting open questions, emerging approaches, and where the field needs to go next.Senior TRAILS researchers come together to reflect on where AI evaluation is headed and what it will take to assess AI systems in ways that actually matter in real-world use. The conversation will span technical, social, and governance perspectives, highlighting open questions, emerging approaches, and where the field needs to go next. ***Panelists****: Susan Ariel Aaronson (GW University); David Broniatowski (GW University); Virginia Byrne (Morgan State University); Hal Daumé III (University of Maryland), Tom Goldstein (University of Maryland); Katie Shilton (University of Maryland)* ***Moderator:*** *Cody Buntain (University of Maryland)*
5:15 – 6:30 p.m.	**Reception and Poster Session**
March 5, 2026	
8:30 – 9 a.m.	**Opening Plenary Session**
9 – 10:15 a.m.	**Interactive Breakout Sessions (Concurrent)** ● **Making Participatory AI Evaluation Fun (A Game Show Simulation)** ○ *Speakers: Katie Shilton (University of Maryland), Jordan Boyd-Graber (University of Maryland)* ● **When Accuracy Isn't Enough: Governing AI by Connecting Metrics to Human and Organizational Outcomes** ○ *Speakers: Stella Umunna (Cantor Fitzgerald; Doctoral*

	Candidate at GW University), George Rivera (Executive Leader and Organizational Change Facilitator) This interactive simulation will explore the tension between technical performance and human consequences, building the discernment skills required for modern AI leadership. To build AI that truly serves society, evaluation must move beyond a one-time technical checkpoint and become a continuous, human-centered practice. This session explores five dimensions of human-centric evaluation: context-aware deployment, human impact and trust, organizational accountability across the AI lifecycle, indicators that link technical signals to social outcomes, and scalable structures for review and oversight. ● **Roundtable: Trustworthiness of AI Research Tools for Literature Reviews** ○ *Facilitator: Larry Liu (Morgan State University)* AI tools are increasingly used to search, summarize, and synthesize research, but their reliability and limits are not always clear. This roundtable explores where AI-assisted literature reviews can add real value and where they may introduce risks, including bias, inaccurate citations, and over-reliance on automated summaries. Participants will discuss how to use these tools responsibly, how to maintain transparency and accountability in research workflows, and how to distinguish helpful assistance from practices that weaken evidence or obscure intellectual effort. The goal is to develop practical, shared guidelines for integrating AI into literature reviews without compromising rigor or trust.
10:15 – 10:45 a.m.	**Networking Break**
10:45 a.m. – 12 p.m.	**Interactive Breakout Sessions (Concurrent)** ● **Transparency as a Tool for Accountability** ○ *Speakers: Susan Ariel Aaronson (GW University); Ilan*

	Strauss (AI Disclosures Project, Social Science Research Council) ● **Evaluating AI for Startups: Practical Metrics for Safety, Adoption, and Competitive Inclusion** ○ *Speakers: Hua Wang (Executive Director, Global Innovation Forum)* AI evaluation frameworks often assume resources and capabilities that small and mid-sized enterprises (SMEs) do not have. This session explores how to design evaluation practices that are scalable, affordable, and meaningful for smaller organizations while maintaining safety and accountability. Participants will learn lightweight evaluation approaches that SMEs can apply across common AI use cases. The session also highlights practical governance and risk safeguards that can be implemented without large compliance teams. Finally, it introduces actionable policy and ecosystem interventions that can help SMEs evaluate AI responsibly. ● **Evaluating LLMs for Clinical Decision Support on the Front Lines** ○ *Speakers: Nirmal Ravi (Atri Consulting), Robert Pless (GW University)* This session presents a real-world evaluation of LLM-based clinical decision support used by frontline health workers in two outpatient clinics in Kano, Nigeria. Community health workers received AI feedback on draft care plans, which were compared to plans created without AI support. While LLM feedback led to noticeable changes in diagnoses, testing, and prescribing—and performed well in chart-based reviews—it did not significantly improve diagnostic accuracy or treatment decisions when compared with on-site physicians' assessments and laboratory results. The case highlights a gap between

	documentation-focused evaluations and clinically meaningful outcomes. Participants will gain practical insights into evaluation design, human–AI interaction, and why common methods may overstate the real-world impact of LLMs in clinical care.
12 – 1 p.m.	**Lunch**
1 – 2:45 p.m.	**GW Trustworthy AI Sandbox Workshop** This workshop will focus on applications of a socio-technical sandbox (or testbed) as both a collaborative environment and integrated toolkit to rapidly prototype and evaluate alternative workflow integrations, task assignments, outcome optimization approaches, and governance strategies. We will begin with a brief overview of the motivations, design principles, and core architecture of the sandbox, followed by a live demonstration. Participants will then explore how TRAILS researchers might leverage the platform to observe, simulate, or experimentally assess social and technical interactions within controlled yet realistic research settings. ***Speakers:*** *Zoe Szajnfarber (GW University), Ryan Watkins (GW University)*
1 – 2:15 p.m.	**Interactive Breakout Sessions (Concurrent)** ● **How To Determine if an AI System is Trustworthy in Practice** ○ *Facilitator: David Broniatowski (GW University)* ● **Responsible AI Governance Crash Course: From Vendor Transparency to Policymaking** ○ *Speakers: Maddy Dwyer (Center for Democracy & Technology), Quinn Anex-Ries (Center for Democracy & Technology)* In this session, experts in state and local AI governance from the Center for Democracy & Technology (CDT) will

	explore their recently developed responsible AI policymaking checklist for elected officials and senior agency leaders and their nine category rubric that helps public administrators assess the transparency of products offered by public sector AI vendors. Both policymaking and procurement of AI products work hand-in-hand to ensure that the benefits of implementing an AI system outweigh the potential risks to constituents. CDT experts will walk participants through how they can promote public transparency and stakeholder engagement; accuracy and reliability; governance and coordination; privacy and security; and safety, rights, and legal compliance in their AI policymaking and how to use the rubric to assess the transparency of current and potential AI tools. ● **From Threats to Vulnerabilities: Assessing Security Risk in Generative AI Systems** ○ *Speakers: Elie Alhajjar (RAND), Kyle Killian (RAND), Sasha Romanosky (CVSS)* This interactive session presents a vulnerability-focused approach to evaluating the security and reliability of generative AI systems, particularly where traditional software assurance methods are limited. Instead of centering on adversary behavior or specific attack techniques, the framework emphasizes identifying and categorizing system weaknesses across areas such as training data, tokenization, context management, and plugin integration. Participants will examine how these vulnerabilities differ from conventional software flaws, why some cannot be fully remediated, and how evaluation practices must adapt to the dynamic nature of generative AI. Through guided exercises and real-world examples, attendees will map vulnerabilities to potential risks and discuss practical approaches for evaluation and mitigation in their own organizational contexts.
2:30 – 3:45 p.m.	**Interactive Breakout Sessions (Concurrent)**

	● Metrics and Methods Roundtable (TBA) ○ Facilitator: Tom Goldstein (University of Maryland) ○ Abstract: Coming soon! ● **The Pause Before AI: Distinguishing When AI Should Lead, Assist, or Stay Out** ○ *Speakers: Sabrina Papazian (Lyft), Elyse Nicolas (GW University)* The year 2025 was about learning how to apply AI into our workflows, understanding when and where it works well. 2026 is the year that cements these practices into permanent habits. Before we lock in those habits, we need to pause and ask ourselves: why are we applying AI to this specific task? Is it a genuine value-add, or an act of avoidance - reaching for AI because it's easier, "expected" of us, or simply there? In industry and beyond, there's been top-down pressure to apply AI to everything. "Just tell the AI to do it" is something we have all heard. But what if relying too heavily on AI, we are losing our ability to think, process, and do good work? This session helps participants distinguish when they are turning to AI for collaboration rather than for abdication. Through collaborative discussions in a workshop setting and scenario testing, participants will co-create a practical set of evaluative questions to determine whether AI should lead, assist, or stay out entirely - questions they can immediately apply in their own work.
3:55 – 4:30 p.m.	**Collective Synthesis: What Did We Just Build Together?** Over the last two days, we have been thinking hard about intersecting hard problems from myriad perspectives. Let's harvest that collective intelligence before we go our separate ways. In this closing session, you'll discuss and share your takeaways and calls to action. We'll then use live AI analysis of your responses to help us ask the most important question of the day: Where do we go from here? Stick

	around and help us find out what we built together. ***Moderator:*** *Darren Cambridge (Managing Director, TRAILS)*