Principled Design for Trustworthy AI:
Interpretability, Robustness, and Safety Across Modalities

1. Workshop Summary

Modern AI systems, particularly large language models, vision-language models, and deep vision networks, are increasingly deployed in high-stakes settings such as healthcare, autonomous driving, and legal decisions. Yet, their lack of transparency, fragility to distributional shifts between train/test environments, and representation misalignment in emerging tasks and data/feature modalities raise serious concerns about their trustworthiness.

This workshop focuses on **developing trustworthy AI systems by principled design**: models that are interpretable, robust, and aligned across the full lifecycle – from training and evaluation to inference-time behavior and deployment. We aim to unify efforts across modalities (language, vision, audio, and time series) and across technical areas spanning interpretability, robustness, uncertainty, safety, and policy.

Our goal is to create a workshop platform for cross-disciplinary discussion and idea exchange across key dimensions of trustworthiness in modern AI systems. These include interpretability & mechanistic transparency, uncertainty quantification & risk assessment for safe operation, adversarial & distributional robustness, and representation & safety alignment across diverse tasks & modalities. By bringing together these efforts under a cohesive design paradigm, the workshop seeks to advance forward-looking solutions and foster community building around shared technical & societal challenges in building trustworthy AI systems. This workshop differs from the recent prior workshop efforts (e.g ICML'24 TiFA, NeurIPS'24 Interpretable AI, IJCAI'24 Trustworthy AI) in its unique focus on building Trustworthy AI systems *by design* and its broad coverage of the *full* machine learning lifecycle across both single- and multi-modal settings.

Topics of interest include 6 pillars:

(1) Interpretable and Intervenable Models: concept bottlenecks and modular architectures, neuron tracing and causal influence methods, mechanistic interpretability and concept-based reasoning, interpretability for control and real-time intervention;

(2) Inference-Time Safety and Monitoring: reasoning trace auditing in LLMs and VLMs, inference-time safeguards and safety mechanisms, chain-of-thought consistency and hallucination detection, real-time monitoring and failure intervention mechanisms;

(3) Multimodal Trust Challenges: grounding failures and cross-modal misalignment, safety in vision-language and deep vision systems, cross-modal alignment and robust multimodal reasoning, trust and uncertainty in video, audio, and time-series models;

(4) Robustness and Threat Models: adversarial attacks and defenses, robustness to distributional, conceptual, and cascading shifts, formal verification methods and safety guarantees, robustness under streaming, online, or low-resource conditions;

(5) Trust Evaluation and Responsible Deployment: human-AI trust calibration, confidence estimation, and uncertainty quantification, metrics for interpretability, alignment, and robustness, transparent, reproducible, and accountable deployment pipelines, safety alignment in fine-tuning, instruction-tuning, and retrieval-augmented systems.

(6) Safety and Trustworthiness in LLM Agents: Autonomous tool use and agentic behavior in LLMs, Safety and failures in planning and action execution, emergent behaviors in multi-agent interactions, intervention and control in agent loops, alignment of long-horizon goals with user intent, auditing and debugging LLM agents in real-world deployment.

The workshop targets a diverse audience from academia, industry, and government, and encourages interactive formats that foster practical insights and collaboration, ultimately paving the way for next-generation AI ecosystems that are more trustworthy, transparent, and conducive to broad public access and democratization. In particular, our workshop will feature a number of invited talks from leading researchers in both academia and industry, selected contribution talks from participants, and poster sessions for accepted papers.

2. Tentative Schedule:

The workshop will last one day, all talks will be planned in person. It will consist of **three** main parts: (1) **invited keynotes** with discussion, (2) **spotlight** contributed talks with discussion, (3) **poster** session *Target Audience*: Researchers, students, and industrial practitioners interested in Trustworthy machine learning, statistics as well as potential applications of deep learning. *Paper Review Process*: We will leverage the existing rich infrastructure that we have built for our previous NeurIPS workshop: a friendly website featuring all details of our workshop, a paper reviewing process built on the OpenReview system with a large diverse pool of reviewers from many different institutions to handle large load of paper submissions as well as to avoid potential conflict of interest. We will put the detailed schedule and talk titles up publicly prior to site publication and note the archival status of the submissions on the website.

- 9:00 am -9:15 am: Opening remarks
- 9:20 am -9:50 am: Invited talk 1 (25 min+5 min Q/A)
- 9:50 am -10:20 am: Invited talk 2 (25 min + 5 min Q/A)
- 10:20 am -10:50 am: Invited talk 3 (25 min + 5 min Q/A)
- 11:00 am -12:00 pm: Poster session & Networking socials
- 12:00 pm - 1:30 pm: Lunch break
- 1:30 pm - 1:45 pm: Spotlight contributed talk 1 (15 min)
- 1:45 pm - 2:00 pm: Spotlight contributed talk 2 (15 min)
- 2:00 pm - 3:00 pm Poster session & Networking socials
- 3:00 pm - 3:30 pm: Invited talk 4 (25 min+5 min Q/A)
- 3:30 pm - 4:00 pm: Invited talk 5 (25 min+5 min Q/A)
- 4:00 pm - 4:30 pm: Invited talk 6 (25 min+5 min Q/A)
- 4:30 pm - 4:45 pm: Spotlight contributed talk 3 (15 min)
- 4:45 pm - 5:00 pm: Spotlight contributed talk 4 (15 min)
- 5:00 pm - 5:15 pm: Concluding remarks

3. Invited speakers

All of our invited speakers have confirmed (5 confirmed, 1 confirmed likely) to give a talk.

Dr. Yan Liu (confirmed), University of Southern California, webpage, area: Interpretability, Robustness
Yan Liu is a full Professor in the Computer Science Department, Director of USC Machine Learning Center, and Co-Director of USC Institute of Ethics and Trust in Computing. Her research focuses on time series modeling and explainable machine learning models with applications to healthcare, sustainability (climate science, traffic), and social media analysis. She has won the best paper award in SIAM conference on Data Mining, NSF CAREER Award, Faculty awards from JP Morgan, Adobe, IBM, and Facebook.

Dr. Mihaela van der Schaar (confirmed), University of Cambridge, webpage, area: Interpretability
Mihaela van der Schaar is a full professor of machine learning, artificial intelligence and medicine at the University of Cambridge. Her current research focuses on interpretable machine learning and machine learning for healthcare. She is an IEEE Fellow and has received numerous awards, including the NSF CAREER Awards, Johann Anton Merck Award, the Oon Prize on Preventative Medicine, 3 IBM Faculty

Awards, and several best paper awards including the IEEE Darlington Award. She was identified by the National Endowment for Science, Technology and the Arts as the most cited female AI researcher in U.K.

Dr. Nanyun (Violet) Peng (confirmed), UCLA, webpage, area: AI alignment and safety
Violet Peng is an associate professor at the Computer Science department of UCLA and an Amazon Scholar at the Amazon AGI org. Her vision is to develop robust NLP techniques to lower communication barriers and make AI agents true companions for humans. Her research focuses on AI alignments including controllable language generation, multi-modal foundations models, and automatic evaluation of foundation models. Her work has been recognized with multiple paper awards, including an Outstanding Paper Award at NAACL 2022, three Outstanding Paper Awards at EMNLP 2024, and Best Paper Awards at AI/NLP workshops. Her research has received support from the NSF CAREER Award, NIH R01, DARPA, IARPA grants, and multiple industrial research awards.

Dr. Hamed Hassani (confirmed), UPenn, webpage, area: Robustness and AI safety
Hamed Hassani is an associate professor in the department of Electrical and System Engineering. He is the site Lead of EnCORE: Institute for Emerging CORE Methods of Data Science. His recent research focuses on AI safety and adversarial robustness of LLMs including jailbreak methods. He has won IEEE Communications Society & Information Theory Society Joint paper award, Intel Rising Star Faculty Awards, NSF CAREER Award, and AFOSR Young Investigator Award.

Dr. Martin Wattenberg (confirmed likely), Harvard University, webpage, area: Interpretability
Martin Wattenberg is a Gordon McKay Professor of Computer Science at Harvard University. His work in machine learning focuses on transparency and interpretability, as part of a broad agenda to improve human and AI interaction. He is well known for his contributions to social and collaborative visualization, and the systems he has created are used daily by millions of people. Martin is also known for visualization-based artwork, which has been exhibited in venues such as the Museum of Modern Art in New York, London Institute of Contemporary Arts, and the Whitney Museum of American Art.

Dr. Fernanda Viegas (confirmed), Harvard University, webpage, area: Interpretability
Fernanda Viegas is a Gordan McKay Professor of Computer Science at Harvard with an affiliation at Harvard Business School. She is also a principal scientist at Google, where she co-founded the PAIR (People + AI Research) initiative and the Big Picture Team. Her research focuses on interpretable machine learning, data visualization, and evaluation of AI systems, which focuses on improving human/AI interaction with a broader agenda of democratizing AI technology. She has also contributed to social and collaborative visualization, where her artwork has been exhibited worldwide.

4. Organizers and Biographies

Lily Weng, Assistant Professor in UC San Diego, lweng@ucsd.edu
Her research focuses on Trustworthy Machine Learning with specialization in mechanistic interpretability, robustness, and AI Safety. She is an expert in Trustworthy Machine Learning and has co-organized several workshops and related experiences, including workshops in NeurIPS'21, AAAI'23, ICCV'25, and GenAI summit'25 in UCSD. Her research expertise and relevant experience in organizing workshops can make the proposed workshop successful. She is not proposing any other workshops for ICLR 2026.

Nghia Hoang, Assistant Professor in Washington State University, trongnghia.hoang@wsu.edu
Nghia Hoang works on probabilistic and distributed machine learning with applications in federated learning, meta learning, and (offline) black-box optimization in science, engineering, and industrial domains. He has co-organized the NeurIPS'21 Workshop on New Frontiers on Federated Learning. He is not proposing any other workshops for ICLR 2026.

Tengfei Ma, Assistant Professor in Stony Brook University, tengfei.ma@stonybrook.edu

Tengfei Ma's research spans machine learning, natural language processing, and biomedical research. He is particularly interested in exploring and modeling relational structures of data, especially the graphs which characterize interactions of a complex system. He also works on improving the interpretability and trustworthiness of machine learning, including exploring the safety of LLMs, machine-generated code detection, and interpretability for time series analysis. He is not proposing any other workshops for ICLR 2026.

Jake Snell, Associate Research Scholar at Princeton University, jsnell@princeton.edu
Jake is in the computer science department at Princeton University. His research focuses on building large-scale deep learning systems that are robust, reliable, and transparent by leveraging the strengths of probabilistic modeling. He carries expertise in meta-learning, uncertainty quantification, interpretable deep learning, and Bayesian models. He has co-organized a weeklong workshop on the fundamentals of machine learning that took place during Princeton's Wintersession 2024. His technical expertise and previous organizing experience will contribute to a successful workshop. He is not proposing any other workshops for ICLR 2026.

Francesco Croce, Assistant Professor at Aalto University, francesco.croce@aalto.fi
His research mainly focuses on multimodal foundation models, in particular their adversarial robustness (jailbreaks, backdoors, etc.) for safe AI systems, and the ability of multimodal models to capture different aspects of human perception, e.g., visual and semantic similarity. He was co-organizer for the robustness workshops "A Blessing in Disguise: The Prospects and Perils of Adversarial Machine Learning" (ICML 21) and "The Art of Robustness: Devil and Angel in Adversarial Machine Learning" (CVPR 22), as well as for the 1st and 2nd Workshops on "Test-Time Adaptation" (CVPR 24 and ICML 25). He is proposing another workshop for ICLR 2026.

Chandan Singh, Microsoft Research, csinva23@gmail.com
Chandan works on interpretable machine learning with the broad goal of improving science and medicine using data. Recently, he has focused on LLMs and how they can be used to directly explain data in language neuroscience. Separately, he has also worked on developing highly accurate transparent models, such as improving linear models and decision trees. He received his PhD from UC Berkeley in 2022. He is not proposing any other workshops for ICLR 2026.

Subarna Tripathi, Principal Engineer, AI Research Science, Intel Corporation, subarna.tripathi@intel.com
Subarna is a computer vision expert and leads a team that makes contributions to foundational computer vision algorithms, multimodal learning, develops applications in video understanding- generation and builds open-source tools to enable the broader researcher and developer community. She graduated with a PhD from UCSD in 2018 and has been featured as a notable URM alumni at UCSD. Subarna had organized workshops at CVPR, ICCV. She is not proposing any other workshops for ICLR 2026.

Lam Nguyen, IBM Research, lamnguyen.mltd@ibm.com
Lam Nguyen is a Staff Research Scientist at IBM Research working in the intersection of Optimization and Machine Learning/Deep Learning. His research interests include design and analysis of learning algorithms, optimization for representation learning, dynamical systems for machine learning, federated learning, reinforcement learning, time series, and trustworthy/explainable AI. He organized several workshops in NeurIPS'21 and AAAI 2023. He is also in the organizing committee for NeurIPS 2023-2025. He is not proposing any other workshops for ICLR 2026.

Program Committee Member. We have recruited a total of 74 program committee members, and among them we have 42 agreed, 2 tentatively agreed, and 30 tentative to serve as reviewers for our workshop if accepted.

5. Anticipated Audience Size

Based on our experience in hosting workshops and broad audience interest in Trustworthy AI, we estimated *the number of attendees* to be 150-200 people in the room and cover 250-300 audiences in total throughout the event.

6. Plan to get an audience for a workshop (advertising, reaching out, etc)

Based on our past experience of organizing workshops, we will attract audience by the following plans:

- We will build an official workshop website to timely announce our workshops with relevant details
- We will broadly advertise our workshop CFP through relevant AI/ML mailing lists, social medials (Linked-In, X, bluesky) in academia, industries, government agencies, and general public
- We will also advertise our workshop at relevant AI/ML conference/venues to encourage attendance

7. Diversity Commitment

We have made efforts to assemble a diverse team of organizers and speakers that reflects the global and multidisciplinary nature of the Trustworthy AI community. Our efforts aim to ensure broad representation across multiple demographic dimensions among organizers and speakers:

- **Geography:** we represent a geographically diverse group across US and Europe
- **Identity:** we represent diverse backgrounds and identities across gender, race, ethnicity
- **Affiliation:** we come from both academia and industry, fostering cross-sector perspective
- **Career stage:** we cover diverse career stages including early-career (postdoc, assistant professors), mid-career (associate professors), and senior researchers (full professors)
- **Disciplinary:** we span a broad range of Trustworthy AI topics, from theoretical foundations to practical applications, across diverse modalities such as vision, language, multi-modal systems.

These efforts reflect our strong commitment to fostering an inclusive environment and amplifying diverse voices across geography, identity, affiliation, career stage, and disciplinary coverage.

Our Program committee also includes a group of 74 reviewers from numerous institutions across diverse geographic regions and areas of expertise, whose collective knowledge will further support our review efforts.

8. Virtual access to workshop materials and outcome

We will provide virtual access to workshop related materials on our workshop website. We plan to use a similar style of our previous NeurIPS workshop website.

9. Previous related workshop

This is a new workshop. This workshop differs from the recent prior workshop efforts (e.g ICML'24 TiFA, NeurIPS'24 Interpretable AI, IJCAI'24 Trustworthy AI) in its unique focus on building Trustworthy AI systems *by design* and its broad coverage of the *full* machine learning lifecycle across both single- and multi-modal settings.