ICLR 2026 WORKSHOP ON AGENTS IN THE WILD: SAFETY, SECURITY, AND BEYOND

1 WORKSHOP SUMMARY

AI agents are rapidly being deployed across critical real-world applications including healthcare, finance, education, transportation, and scientific research [1–5]. These systems can autonomously interact with the world, make consequential decisions, and execute actions with irreversible real-world consequences. Recent incidents underscore the urgency of addressing agent safety and security: cases involving AI-related harm range from chatbot interactions associated with mental health crises to autonomous systems making critical errors in high-stakes domains [6–8]. International bodies have called for rigorous frameworks to ensure safe agent deployment, as highlighted in reports such as the International AI Safety Report [6, 9, 10]. Despite this urgency, the safety and security of AI agents lack principled theories and comprehensive technical solutions. Moreover, addressing these challenges requires interdisciplinary stakeholders spanning computer science, security, ethics, policy, law, healthcare, education, and domain expertise across different cultures and languages. Currently, there is no dedicated venue that brings together this diverse community to focus specifically on the safety, security, and trustworthiness of AI agents deployed in the wild. Our workshop aims to discuss and approach the fundamental challenges in agents in real-world deployment settings or "agents in the wild":

Challenge 1: Foundational Gaps in Principled Safety, Security, and Trustworthiness for Agentic AI. Existing safety, security, and trustworthiness principles developed for traditional software systems or even for foundation models do not trivially transfer to agentic settings. Unlike static foundation models that generate text in controlled environments, agents take actions, use external tools, maintain state across extended interactions, and operate autonomously in dynamic environments [3, 4]. For example, while adversarial robustness in foundation models focuses on input perturbations affecting outputs, agents face compounding challenges where errors can cascade through multi-step reasoning chains, tool executions, and environmental interactions [11, 12]. This differs fundamentally from traditional software security, where bugs or backdoors affect program execution in predictable ways, whereas agent vulnerabilities can manifest through subtle misalignments in goal-directed behavior that only become apparent after irreversible actions.

Challenge 2: Amplification of Existing Risks and Emergence of Novel Agent-Specific Threats. Agents amplify existing safety and security problems while introducing entirely new risks [13, 14]. Methods that successfully align foundation models can fail catastrophically when those models operate as agents with tool access [15, 16]. For instance, models that reliably write secure code may become susceptible to inserting backdoors when operating autonomously with the ability to modify codebases [15, 17, 18]. For example, security challenges specific to agents, such as prompt injection attacks that hijack agent behavior [13, 19], malicious tool use, and adversarial manipulation of multi-agent coordination, have not been extensively discussed in existing literature. Privacy violations can occur when agents inadvertently leak sensitive information through tool calls or cross-context interactions. Hallucinations become more dangerous when agents act on false information rather than merely presenting it [20, 21].

Challenge 3: Asymmetric Defense Challenges in Open-Ended Agentic Systems. Defense is inherently harder than attack in agentic systems. For example, an agent equipped with advanced reasoning capabilities could potentially identify and exploit zero-day vulnerabilities more efficiently than current automated tools, creating an asymmetry where malicious actors gain disproportionate advantages [19, 22]. As agentic autonomy grows, human-in-the-loop verification at critical decision points becomes increasingly infeasible, allowing problems to remain hidden until after irreversible consequences occur [2]. The attack surface expands dramatically. For example, beyond traditional software vulnerabilities, agents introduce risks through prompt injection, tool misuse, goal misalignment, and coordination failures in multi-agent settings [13, 23, 24].

Challenge 4: Absence of Comprehensive and Realistic Evaluation Frameworks for Agent Safety. There is a critical lack of comprehensive evaluation benchmarks that assess agent safety

and security in realistic deployment scenarios [14, 22]. Existing benchmarks primarily measure capability rather than safety, and often evaluate agents in simplified settings that fail to capture the complexity and adversarial nature of real-world deployments. Literature focusing on agentic misalignment in the wild [14] utilizes non-standardized tasks in an artificial setting to force binary decision making. Standardized frameworks are needed to rigorously assess whether agents are ready for deployment in high-stakes domains.

Challenge 5: Rapid Evolution and Increasing Complexity of Agentic Ecosystems. Agent systems are rapidly evolving in ways that make safeguarding increasingly difficult, continually expanding the threat frontier and rendering static defense strategies ineffective. For example, multi-agent systems introduce emergent behaviors and coordination vulnerabilities that do not exist in single-agent settings [11, 25, 26]. Multi-turn interactions over extended time horizons create opportunities for subtle goal drift and accumulating errors. Multimodal and computer-use agents that integrate vision, language, and physical action face safety challenges unique to their environments [19]. Post-training processes like fine-tuning and reinforcement learning can inadvertently undermine safety properties established during initial training [27–29]. Each of these developments introduces new attack surfaces and failure modes that must be understood and addressed.

Topics of Interest: In response to these challenges, our workshop will explore, but not be limited to, the following topics:

- **Agentic Safety and Alignment:** How can we ensure agents are not used to perform harmful actions (e.g., the creation of weapons of mass destruction)? How can we ensure the safety of agentic tool use? How can we ensure agents remain aligned with human values in long-term and adversarial settings?
- **Agent Security, Privacy, and Robustness:** How should we evaluate and assess the emergence of new attack surfaces and vulnerabilities unique to agentic systems and their implications for security practice (e.g., redteaming agents)?
- **Agentic Hallucination and Factuality:** How can we detect and mitigate hallucinations in agent outputs, especially when agents use tools and retrieve external information?
- **Agentic Interpretability and Transparency:** How can we audit, explain, and steer agent alignment behaviors and ensure agents are not deceiving us?
- **Agentic Fairness and Bias:** How can we ensure agents make fair and equitable decisions across diverse populations, settings, and contexts?
- **Evaluating and Benchmarking Agents:** How can we construct benchmarks to accurately assess real-world agentic risks?
- **Multimodal and Computer-Use Agents:** How can we explore safety challenges unique to agents that integrate multiple modalities?
- **Multi-Agent Coordination and Long-Horizon Safety:** How should we approach and assess agents in different multi-agent and long-horizon settings?
- **Post-Training and Adaptating Agents:** How do post-training methods impact model safety and alignment in agentic settings?
- **Agent Systems and Infrastructure:** How do we enable scalable and efficient agent deployments that still allow for human monitoring and consistent safety checking?
- **Interdisciplinary Agentic Considerations:** How do we address cross-disciplinary challenges in agent safety spanning cultural, legal, environmental, and domain-specific dimensions?
- **Ethics, Society, and Governing of Agents:** How should agentic systems be regulated, governed, and ethically used in social, environmental, and legislative contexts?

We particularly encourage submissions that bridge multiple areas, offer novel perspectives on emerging challenges, present lessons learned from deployment experiences, or propose ambitious research agendas for the field.

This workshop aims to create a venue for researchers and practitioners to discuss and establish the foundations of safe, secure, and trustworthy agentic AI. Our goal is to assemble an interdisciplinary

effort to identify the most significant open problems, propose concrete research directions, and chart a roadmap for advancing the long-term aligned deployment of agentic systems.

2 TENTATIVE SCHEDULE & PLANS

The workshop will be a full-day in-person event designed to maximize opportunities for knowledge exchange and community building. While we encourage in-person attendance, we plan to provide options for online participation. The workshop will feature eight 30-min invited talks (25-min talk + 5-min Q&A) from distinguished speakers representing diverse perspectives from academia and industry, and six 15-min spotlight presentations (10-min talk + 5-min Q&A). Four of the spotlight presentations will be selected from the regular papers track, and two will be selected from the tiny papers track. Moreover, the workshop will include two 1-hour poster sessions for accepted papers, and a 45-min panel discussion, addressing future directions and open challenges for developing and deploying safe and secure agentic AI. Communication between all attendees will be mediated via a dedicated channel hosted on Rocket.Chat.

2.1 WORKSHOP SCHEDULE

Morning Session:

- 08:20 - 08:30: Opening Remarks
- 08:30 - 09:00: Invited Talk 1
- 09:00 - 09:30: Invited Talk 2
- 09:30 - 10:30: Poster Session 1
- 10:30 - 11:00: Invited Talk 3
- 11:00 - 11:15: Spotlight Presentation 1
- 11:15 - 11:30: Spotlight Presentation 2
- 11:30 - 12:00: Invited Talk 4
- 12:00 - 12:30: Invited Talk 5
- 12:30 - 13:30: Lunch Break

Afternoon Session:

- 13:30 - 14:00: Invited Talk 6
- 14:00 - 14:15: Spotlight Presentation 3
- 14:15 - 14:30: Spotlight Presentation 4
- 14:30 - 15:30: Poster Session 2
- 15:30 - 16:00: Invited Talk 7
- 16:00 - 16:15: Spotlight Presentation 5
- 16:15 - 16:30: Spotlight Presentation 6
- 16:30 - 17:15: Panel Discussion
- 17:15 - 17:45: Invited Talk 8
- 17:45 - 18:00: Awards and Closing Remarks

The workshop will start at 8:20 and end at 18:00 to ensure a comfortable and balanced schedule for all attendees. We alternate the talks (invited talks, spotlight presentations) and open discussions (poster sessions, panel discussion) to encourage discussion and provide opportunities for attendees and speakers to connect. Moreover, we will post details of the talks and posters on the workshop website prior to the workshop to allow attendees to choose which sessions to attend and enable cross-participation between different workshops.

2.2 PAPER SUBMISSION AND REVIEW PROCESS

The workshop will solicit two types of submissions to accommodate diverse contributions from the research community:

- **Regular Papers Track:** The workshop welcomes submissions of research and position papers, with options for both long (9-page) and short (4-page) submissions, related to developing and deploying AI agents in the wild. References and supplementary materials will not count against these limits.
- **Tiny Papers Track:** To encourage submissions from underrepresented and under-resourced researchers, we will incorporate a tiny papers track with a 2-page limit. These submissions can present recent breakthroughs, unpublished ideas, modest theoretical results, follow-up experiments, or fresh perspectives on existing work.

All submissions must be in a single PDF file following the ICLR 2026 template and submitted through the OpenReview submission portal.

Important Dates. We will follow the suggested dates by ICLR for paper submission and review:

- **Paper Submission Open:** January 1, 2026
- **Paper Submission Deadline:** January 30, 2026 AoE

- **Paper Notification:** March 1, 2026 AoE
- **Camera-ready Version Deadline:** March 10, 2026 AoE
- **Workshop Date:** April 26 or 27, 2026

Review Process. All submissions will undergo double-blind peer review, with each submission receiving at least 3 reviews. We will reach out to a diverse pool of reviewers, including the authors of submitted works, and ensure a maximum of 4 assigned papers per reviewer. We will use OpenReview for the review process to ensure transparency and enable constructive feedback from the broader community. The reviewers will be instructed to evaluate submissions based on technical quality, relevance to the workshop themes, clarity of presentation, and potential to generate interesting discussions. Reviewers will be given the opportunity to flag a submission for ethics review or to nominate a submission for spotlight presentation or award via OpenReview. Decisions regarding acceptance, spotlight presentations, ethical reviews, and awards will be made transparently by the organizing committee.

Conflicts of Interest. The organizing committee will proactively search for any conflicts of interest using OpenReview profiles of authors and reviewers. We will ensure that reviewers are not assigned any submissions from their organization. Similarly, members of the organizing committee will not be involved in the assessment (e.g., acceptance decisions, ethics review, spotlight presentations, and awards) of any submission from the same organization.

LLM Usage Policy. The workshop will follow ICLR's policies on LLM usage. To ensure the guidelines are consistent across the submission tracks, we will not allow AI-generated papers. Of course, AI assistance is permitted, but submissions must be primarily human-authored, reflecting original thought and analysis.

Accepted Papers. The workshop will adopt a non-archival policy, welcoming ongoing and unpublished work, as well as papers under review or recently accepted at other venues (provided they do not breach dual-submission or anonymity policies). We will discourage the submission of works previously published at major venues (e.g., ICLR, ICML, NeurIPS). This policy enables broader participation and discussion of cutting-edge research, while keeping the focus on submissions that would most benefit from feedback at ICLR.

All accepted papers will be designed for poster presentations, and a select few (4 from the regular track and 2 from the tiny track) will be additionally selected for spotlight presentations. Accepted papers will be published on the workshop website and OpenReview, making them accessible to the broader research community. Additionally, we will encourage the authors to upload the posters and a brief video overview of their work to be hosted on the workshop website and viewable by remote attendees.

3 Invited Speakers and Panelists

We have invited a total of 9 distinguished and diverse speakers representing major research labs, industry leaders, and top academic institutions. Each speaker brings unique expertise and perspectives on current developments in agent safety, security, trustworthiness, and deployment. This ensures broad coverage of the workshop themes and brings together viewpoints from different research communities and application domains. Confirmed speakers include:

1. **Prof. Yoshua Bengio** (Mila & LawZero & Université de Montréal) **[confirmed]**
    - Talk title: *Superintelligent Agents Pose Catastrophic Risks: Can Scientist AI Offer a Safer Path?*
    - Bio: Yoshua Bengio is Full Professor of Computer Science at Université de Montréal, Co-President and Scientific Director of LawZero, as well as the Founder and Scientific Advisor of Mila and a Canada CIFAR AI Chair. Considered one of the world's leaders in Artificial Intelligence and Deep Learning, he is the recipient of the 2018 A.M. Turing Award, considered to be the "Nobel Prize of computing." He is the most cited computer scientist worldwide, and the most-cited living scientist across all fields (by

total citations). Professor Bengio is a Fellow of both the Royal Society of London and Canada, an Officer of the Order of Canada, a Knight of the Legion of Honor of France, a member of the UN's Scientific Advisory Board for Independent Advice on Breakthroughs in Science and Technology, and chairs the International AI Safety Report.

2. **Dr. Jared Quincy Davis** (Mithril & Stanford University) **[confirmed]**
    - Talk title: *Compound AI Systems: Design Patterns for Secure Multi-Agent Deployments*
    - Bio: Jared Quincy Davis is the Founder and CEO of Mithril, a company orchestrating global compute capacity for AI workloads. He completed his PhD in Computer Science at Stanford University, where his research focused on compound AI systems and distributed machine learning infrastructure. His pioneering work on Networks of Networks addresses fundamental challenges in building reliable multi-agent systems that can scale to production environments. Prior to founding Mithril, he conducted research on large-scale ML systems and contributed to understanding how to compose multiple AI models into robust, scalable applications. His expertise in systems design, distributed computing, and large-scale ML infrastructure provides crucial insights into the practical challenges of deploying agentic AI systems at scale, particularly regarding reliability, fault tolerance, and coordination across heterogeneous computing resources.
3. **Dr. Dan Hendrycks** (Center for AI Safety) **[confirmed]**
    - Talk title: *Measurements For Capabilities And Hazards*
    - Bio: Dr. Dan Hendrycks is the director of the Center for AI Safety and an advisor to xAI and Scale AI. He received his PhD from UC Berkeley where he was advised by Dawn Song and Jacob Steinhardt, and his BS from the University of Chicago. His research is supported by the NSF GRFP and the Open Philanthropy AI Fellowship. He has made foundational contributions to machine learning and AI safety, including the GELU activation function (the most-used activation in state-of-the-art models including BERT, GPT, and Vision Transformers), the out-of-distribution detection baseline, and distribution shift benchmarks. Dr. Hendrycks developed influential benchmarks for AI safety evaluation, including MMLU (Massive Multitask Language Understanding), the Weapons of Mass Destruction Proxy (WMDP) benchmark for measuring hazardous knowledge in biosecurity, cybersecurity, and chemical security, and Humanity's Last Exam in collaboration with Scale AI. He is the author of "Introduction to AI Safety, Ethics, and Society", a comprehensive textbook on understanding AI risk. His work has been featured in major publications including the Wall Street Journal and TIME Magazine.
4. **Prof. Bo Li** (University of Illinois Urbana-Champaign & Virtue AI) **[confirmed]**
    - Talk title: *Guarding the Future: Advancing Risk Assessment, Safety Alignment, and Guardrail Systems for AI Agents*
    - Bio: Dr. Bo Li is an Associate Professor in the Department of Computer Science at the University of Illinois at Urbana-Champaign. She is the recipient of the IJCAI Computers and Thought Award, Alfred P. Sloan Research Fellowship, IEEE AI's 10 to Watch, NSF CAREER Award, MIT Technology Review TR-35 Award, Dean's Award for Excellence in Research, C.W. Gear Outstanding Faculty Award, Intel Rising Star Award, Symantec Research Labs Fellowship, Rising Star Award, Research Awards from Tech companies such as Amazon, Meta, Google, Intel, IBM, and eBay, JPMC, Oracle, and best paper awards at several top machine learning and security conferences. Her research focuses on both theoretical and practical aspects of trustworthy machine learning, which is at the intersection of machine learning, security, privacy, and game theory. Her work has been featured by several major publications and media outlets, including Nature, Wired, Fortune, and New York Times. Her research group has developed influential frameworks for evaluating and improving the robustness of AI agents in real-world settings.
5. **Dr. Yunyao Li** (Adobe) **[confirmed]**

- Talk title: *Building & Querying Enterprise Knowledge Bases: From Declarative Languages to Secure Agents*
- Bio: Yunyao Li is a Director of Machine Learning at Adobe Experience Platform, where she leads strategic initiatives to bring the power of Generative AI and Knowledge Graph to enterprise systems. Previously, she was the Head of Machine Learning at Apple Knowledge Platform. Before joining Apple, she was a Distinguished Research Staff Member and Senior Research Manager at IBM Research - Almaden. Yunyao is an ACM Distinguished Member. She is particularly known for her work in enterprise natural language processing, enterprise search, and database usability. In these areas, she has published over 100 peer-reviewed articles, been granted 36 patents, created and taught multiple graduate-level courses, and co-authored two books. She is an ACM Distinguished Member, a member of the inaugural New Voices program of the American National Academies, and a young scientists at World Laureates Forum Young Scientists Forum in 2019. She was a member of NAACL Executive Board from 2022-2024. She received her undergraduate degrees from Tsinghua University, and her master's and Ph.D. degrees from the University of Michigan - Ann Arbor. Her expertise provides crucial insights into the practical challenges and safety considerations of deploying agentic systems in production environments, particularly around data security, privacy protection, and ensuring reliable operation in enterprise contexts where mistakes can have significant business consequences.

6. **Dr. Bing Liu** (Scale AI) **[confirmed]**
    - Talk title: *From LLMs to Agents: The Evaluation Challenge*
    - Bio: Bing Liu is the Head of Research at Scale AI and Adjunct Professor of Computer Science and Engineering at the University of California, Santa Cruz. His research focuses on LLM post-training, evaluation, and agentic systems. At Scale AI, his team develops open benchmarks and evaluation methodologies, including Humanity's Last Exam and SWE-Bench Pro, to measure and advance the capabilities of frontier models. Prior to Scale, he worked on Llama3 post-training and evaluation at Meta, and led the development of Meta AI Assistant. He previously held research positions at Google Research, and earned his Ph.D. in Electrical and Computer Engineering from Carnegie Mellon University. Bing has published over 40 peer-reviewed papers, and has served as Area Chair and Senior Program Committee member for leading AI and NLP conferences including ICLR, ACL, and EMNLP. His expertise in evaluation methodologies addresses the fundamental challenge of rigorously assessing agent safety, reliability, and robustness across diverse deployment scenarios, essential for understanding when agents are ready for real-world use.
7. **Prof. Yang Liu** (UC Santa Cruz) **[confirmed]**
    - Bio: Yang Liu is an Associate Professor in the Department of Computer Science and Engineering at UC Santa Cruz. He received his PhD from the University of Michigan and was a postdoctoral fellow at Harvard University. His research focuses on trustworthy machine learning, with particular expertise in machine unlearning, fairness in AI, and data-centric approaches to AI safety. He has received multiple awards including the NSF CAREER Award and best paper awards at top venues. His work on machine unlearning addresses how to safely remove information from trained models, which is crucial for agent systems that may need to forget sensitive information or correct learned behaviors. Prof. Liu will serve as the panel moderator, bringing his expertise in trustworthy machine learning to facilitate discussion on open challenges and future directions for safe and secure agent deployment in the wild.
8. **Prof. Yizhou Sun** (UCLA & Amazon) **[confirmed]**
    - Talk title: *Multi-Agent System: From Architecture Design to Real-World Deployment*
    - Bio: Yizhou Sun is a Professor in the Department of Computer Science at UCLA and an Amazon Scholar. She received her Ph.D. in Computer Science from the University of Illinois at Urbana-Champaign in 2012. Her principal research interest is on mining graphs/networks, and more generally in data mining, machine learning, and network science, with a focus on modeling novel problems and proposing scalable algorithms for large-scale, real-world applications. She is a pioneer researcher in mining heterogeneous information network, with a recent focus on deep learning on graphs, AI for

Chip Design, AI for Science, and neuro-symbolic reasoning. Yizhou has over 200 publications in books, journals, and major conferences. Tutorials of her research have been given in many premier conferences. She is a recipient of multiple Best Paper Awards, ACM SIGKDD Doctoral Dissertation Award, Yahoo ACE (Academic Career Enhancement) Award, NSF CAREER Award, CS@ILLINOIS Distinguished Educator Award, Amazon Research Awards (twice), Okawa Foundation Research Award, VLDB Test of Time Award, WSDM Test of Time Award, ACM Distinguished Member, IEEE AI's 10 to Watch, and SDM/IBM faculty award. She is a general co-chair of SIGKDD 2023, PC co-chair of ICLR 2024, and PC co-chair of SIGKDD 2025. Her research on graph-based learning and network analysis provides foundational insights into how information propagates through multi-agent systems and how to detect and prevent cascading failures or adversarial manipulation in distributed agent networks.

9. **Dr. Chi Wang** (Google DeepMind & AG2) **[confirmed]**
    - Talk title: *Frontiers of Agentic AI*
    - Bio: Chi Wang is a Senior Staff Research Scientist at Google DeepMind and the creator and lead developer of AutoGen and AG2, two of the most widely-used open-source frameworks for building multi-agent AI systems. He received his PhD in Computer Science from UIUC. Prior to joining Google DeepMind, he was a Principal Researcher at Microsoft Research, where he led the development of AutoGen, which has been adopted by thousands of developers and researchers worldwide for building agentic applications. His research focuses on the foundations of agentic AI, including agent architectures, multi-agent coordination, tool use, and reasoning capabilities. His work combines theoretical advances in AI with practical system-building experience, addressing scalability, reliability, and safety challenges in deploying multi-agent systems at scale. The AutoGen and AG2 frameworks have become foundational tools for the research community, enabling rapid prototyping and deployment of complex agentic systems while incorporating safety considerations from the ground up.

4 Organizers and Biographies

The organizing committee brings together complementary expertise spanning agent systems, safety, security, multimodal AI, and evaluation. The team includes established faculty with strong track records of organizing successful workshops and early-career researchers who bring fresh perspectives and energy to the community.

Dawn Song (UC Berkeley)

- Email: dawnsong@berkeley.edu
- Webpage: https://dawnsong.io/
- Google Scholar: https://scholar.google.com/citations?user=84WzBlYAAAAJ
- Bio: Dawn Song is a Professor in Computer Science at UC Berkeley and Co-Director of Berkeley Center for Responsible Decentralized Intelligence. Her research interest lies in AI safety and security, agentic AI, deep learning, security and privacy, and decentralization technology. She is the recipient of numerous awards including the MacArthur Fellowship, the Guggenheim Fellowship, the NSF CAREER Award, the Alfred P. Sloan Research Fellowship, the MIT Technology Review TR-35 Award, ACM SIGSAC Outstanding Innovation Award, and more than 10 Test-of-Time Awards and Best Paper Awards from top conferences in Computer Security and Deep Learning. She has been recognized as Most Influential Scholar (AMiner Award), for being the most cited scholar in computer security. She is an ACM Fellow and an IEEE Fellow, and an Elected Member of American Academy of Arts and Sciences. She obtained her Ph.D. degree from UC Berkeley. She is also a serial entrepreneur and has been named on the Female Founder 100 List by Inc. and Wired25 List of Innovators. Her expertise in security and AI safety brings critical insights to the challenges of deploying agents in the wild.

Chenguang Wang (UC Santa Cruz)

- Email: chenguangwang@ucsc.edu

- Webpage: https://cgraywang.github.io/
- Google Scholar: https://scholar.google.com/citations?user=hsZ2aj0AAAAJ
- Bio: Chenguang Wang is an Assistant Professor at UC Santa Cruz specializing in trustworthy large language models and agentic AI systems. He earned a Ph.D. from Peking University, was a postdoctoral researcher at UC Berkeley, and a research scientist at Amazon AI. His research focuses on safety, security, and alignment of AI agents, with applications spanning reasoning, tool use, and multi-agent coordination. He has developed and contributed to impactful open-source systems including rLLM, MassGen, and AutoGluon, and received a 2024 Google Research Scholar Award. His work has been featured in outlets including MIT Technology Review.

Nicholas Crispino (UC Santa Cruz)

- Email: ncrispino@ucsc.edu
- Webpage: https://ncrispino.github.io/
- Google Scholar: https://scholar.google.com/citations?user=oSD_dcgAAAAJ
- Bio: Nicholas Crispino is a second-year PhD student at UC Santa Cruz working on agentic AI systems and multi-agent coordination. His research investigates how multiple agents can cooperate safely and effectively in dynamic environments, addressing challenges in communication protocols, emergent behaviors, and robust coordination strategies.

Ruoxi Jia (Virginia Tech)

- Email: ruoxijia@vt.edu
- Webpage: https://ruoxijia.net/
- Google Scholar: https://scholar.google.com/citations?user=JCrug-YAAAAJ
- Bio: Ruoxi Jia is an Assistant Professor in the Bradley Department of Electrical and Computer Engineering at Virginia Tech. Her research interests span machine learning, security, privacy, and cyber-physical systems, with a recent focus on data-centric and trustworthy AI. Her work has earned her several prestigious awards and fellowships, including the NSF CAREER Award and the Best Social Impact Paper Award at ACL. Her research has been featured in prominent media outlets such as The New York Times, IEEE Spectrum, and MIT Technology Review.

Kyle Montgomery (UC Santa Cruz)

- Email: kylemontgomery@ucsc.edu
- Webpage: https://kylemontgomery1.github.io/
- Google Scholar: https://scholar.google.com/citations?user=O8tnCagAAAAJ
- Bio: Kyle Montgomery is a second-year PhD student at UC Santa Cruz working on LLM post-training, agentic AI, and scaling test-time compute for complex reasoning tasks. His research focuses on methods to improve agent capabilities while maintaining alignment during fine-tuning and evaluation of reasoning processes.

Yujin Potter (UC Berkeley)

- Email: ypotter@berkeley.edu
- Webpage: https://www.linkedin.com/in/yujin-potter-b8a794240
- Google Scholar: https://scholar.google.com/citations?user=ZDG9RD8AAAAJ
- Bio: Yujin Potter is a postdoctoral researcher at UC Berkeley specializing in AI alignment and safety. Her research addresses multi-agent systems, AI bias/fairness, and societal impacts of AI. Before turning to AI safety, she investigated the security of decentralized technologies such as blockchains, DAOs, and DeFi systems.

Vincent Siu (UC Santa Cruz)

- Email: vincent.siu@ucsc.edu
- Webpage: https://vsiu-ucsc.github.io/
- Google Scholar: https://scholar.google.com/citations?user=EiaoeIUAAAAJ
- Bio: Vincent Siu is a first-year PhD student at UC Santa Cruz specializing in mechanistic interpretability and LLM safety. His research explores how alignment features and decision-making processes are represented in large language models, with the goal of developing techniques to better understand and control agent behavior. His research aims to audit and steer agent behavior and ensure transparency in deployed systems.

Zhun Wang (UC Berkeley)

- Email: zhun.wang@berkeley.edu
- Webpage: https://www.linkedin.com/in/zhun-wang-4b7718330/
- Google Scholar: https://scholar.google.com/citations?user=JG_3xhEAAAAJ
- Bio: Zhun Wang is a third-year PhD student at UC Berkeley focusing on secure AI systems and adversarial robustness. His research addresses security vulnerabilities in machine learning models, with particular emphasis on protecting agent systems from adversarial attacks, prompt injection, and other malicious manipulation attempts. His work on red-teaming and defense mechanisms ensures agents can operate safely in adversarial environments.

Workshop Organization Experience. Several of our senior organizers have extensive experience organizing successful workshops at top-tier machine learning conferences. For example, Prof. Dawn Song has organized multiple workshops at ICLR, including the Workshop on Large Language Models for Agents (ICLR 2024), which attracted significant interest in the emerging field of LLM-based agents and featured groundbreaking research on autonomous systems. Prof. Ruoxi Jia and Prof. Dawn Song recently co-organized the Data Problems in Foundation Models workshops at ICLR 2025, addressing critical challenges in data curation, attribution, and quality for foundation models. Additionally, Prof. Ruoxi Jia has delivered tutorials at NeurIPS 2024 on data-centric approaches to foundation model development, including "Advancing Data Selection for Foundation Models: From Heuristics to Principled Methods." Prof. Dawn Song has organized influential massive open online courses (MOOCs) reaching tens of thousands of learners worldwide. Notable examples include the Agentic AI MOOC (Fall 2024, 2025) at UC Berkeley, which most recently attracted over 32,000 registered learners and covered topics spanning agent architectures, reasoning, planning, safety, and real-world deployment. Additional MOOCs have addressed topics including decentralized finance and zero-knowledge proofs, demonstrating a strong commitment to broad educational outreach in emerging AI and technology domains.

Several organizing committee members are first-time workshop organizers (Nicholas Crispino, Kyle Montgomery, Vincent Siu, and Zhun Wang), who bring fresh perspectives and energy to the workshop while being mentored by the experienced senior organizers. This combination of seasoned leadership and emerging talent ensures both the professional execution of the workshop and the cultivation of the next generation of workshop organizers in the community.

5 Workshop Logistics and Virtual Access

Workshop Website. We have created a website to advertise and disseminate the workshop's information, linked here: https://agentwild-workshop.github.io/. We will also use this website to share workshop contributions, including accepted papers, and support future engagement.

Communication. We have set up a Google Group agentwild-workshop@googlegroups.com monitored by the organizing committee to ensure prompt responses to external inquiries. Before the workshop, we will set up a dedicated channel on Rocket.Chat to facilitate interactions among workshop attendants.

Virtual Access. The workshop will primarily be in-person, with robust support for online participation. Each invited talk, spotlight presentation, and panel discussion will be live-streamed via Zoom with real-time Q&A capabilities. All recorded content will be uploaded to the workshop website within 48 hours of the event. We will use Rocket.Chat to collect and moderate questions from remote participants, ensuring they can engage meaningfully with speakers. Authors of accepted works are encouraged to submit posters and 5-minute video overviews, which will be made available on the workshop website one week prior to the event to facilitate asynchronous engagement.

6 Audience Size and Dissemination

Anticipated Audience Size. Based on the popularity of agents in recent literature, broad scope, and audience of previous ICLR workshops, we envision attracting 200-400 submissions and approximately 500 in-person attendees. We invite both theoretical and empirical work from a range of disciplines and encourage interdisciplinary submissions bridging fields like machine learning, robotics, NLP, HCI, ethics/policy, and cognitive science. Moreover, we encourage submissions and attendance from researchers outside the ML conference publication circuit in order to facilitate interdisciplinary discussion and collaboration on agentic AI.

Dissemination. To maximize engagement, the organizing committee plans to dedicate significant effort to promoting this workshop. We've created a workshop website with all relevant information. Before the submission portal opens, we will distribute a call for participation across a variety of ML-related mailing lists, including mailing lists for underrepresented groups in AI, and social media, like X (Twitter) and LinkedIn. The organizers will also advertise the workshop at upcoming ML-related conferences, among research groups, and through our professional networks, in order to reach audiences from both academia and industry.

7 Diversity Commitment

Organizers and Speakers. We aim to promote diversity in all its forms with our selection of organizers and speakers. The organizing committee comprises individuals from varied genders, races, countries, affiliations, and career stages, including PhD students, early-career professors, and established researchers. Several members of the organizing committee are first-time workshop organizers. Our invited speaker lineup similarly reflects diverse perspectives across gender, geographic location, institutional affiliation, and areas of expertise within agent safety and security, ensuring a broad range of topics.

Registration Fee & Travel Grants for Junior Researchers, Local Brazilians, and Underrepresented Groups. To make our workshop more accessible, we aim to offer grants for registration fees and travel expenses to participants who may face financial barriers to attending. Priority will be given to junior researchers (including undergraduate and graduate students, postdocs, and early-career faculty), local Brazilians, and individuals from historically underrepresented groups. We are seeking sponsorship from companies such as Adobe, AG2, Amazon, CAIS, Google DeepMind, LawZero, Mithril, Scale AI, and Virtue AI to support these initiatives and encourage broader participation. Contingent on funding support, we will consider expanding eligibility to participants from nearby South American countries as well.

Participants and Attendees. To enhance the accessibility of our workshop to a broader audience, we encourage authors from other disciplines or outside of the ML conference circuit to submit their work. Moreover, we aim for the tiny papers track to attract submissions from underrepresented or under-resourced researchers. We plan to seek out sponsorships from leading companies in order to offer final support to participants who might otherwise be unable to attend or present due to financial reasons.

Review Process. Our review process for submitted papers will be double-blind (conducted via OpenReview) to mitigate institutional and author biases. The reviewer pool will be curated to ensure broad representation of research areas. Consequently, our workshop will feature a diverse cohort of participants, with selected contributors presenting alongside invited speakers.

Accessibility. All presentation materials and venue arrangements will be designed with accessibility in mind. For those unable to join in-person, we will offer support for online or asynchronous participation. We will adopt ICLR's code of conduct to maintain a welcoming and respectful atmosphere throughout the workshop. As detailed above in the "Virtual Access" section, we plan to offer a wide variety of mechanisms to access or virtually attend the workshop, ensuring equitable access to presentations, talks, and other educational materials presented during the workshop. We believe that diverse perspectives are essential for addressing the multifaceted challenges of agent safety and security, and we are committed to creating an environment where all participants can contribute meaningfully to the discussions and advance the field together.

8 Previous Related Workshops

Our workshop addresses a critical gap in the current workshop landscape by being the first to comprehensively focus on the intersection of agency, safety, security, and trustworthiness for AI agents deployed in real-world environments. Unlike previous workshops that have primarily focused on advancing agent capabilities or treating safety as a separate concern, our workshop explicitly integrates perspectives from safety, security, interpretability, privacy, fairness, and other trustworthiness dimensions from the outset. We are the first workshop to emphasize security challenges specific to agentic systems, including prompt injection, tool misuse, and adversarial attacks on multi-agent coordination, alongside broader trustworthiness concerns such as hallucination, privacy violations, and fairness issues that arise uniquely in agentic contexts. The emphasis on "agents in the wild" reflects the urgent need to address practical deployment challenges as these systems move from research prototypes to production applications with real-world consequences.

This workshop builds on several successful previous events while addressing emerging challenges specific to agentic AI. The Workshop on Large Language Models for Agents (ICLR 2024) explored fundamental capabilities of LLM-based agents, focusing on methods for reasoning, planning, and tool use, while the AIA Workshop (COLM 2025) has examined AI agents more broadly across various applications. The Foundation Models in the Wild workshops (ICML 2024, ICLR 2025) addressed challenges in real-world deployment of foundation models, focusing on adaptation, reliability, and efficiency. Various workshops at NeurIPS and ICLR have addressed AI safety and trustworthy AI from different perspectives, including workshops on Reliable and Responsible Foundation Models (ICLR 2024), Safe Generative AI (NeurIPS 2024), and Next Generation of AI Safety (ICML 2024). While these workshops have made important contributions, advancing our understanding of agent capabilities, foundation model deployment challenges, and general AI safety principles, they have primarily focused on improving performance, addressed abstract or theoretical safety considerations, or examined foundation models without specifically considering the unique challenges of autonomous agents. None have systematically examined the specific security vulnerabilities and trustworthiness challenges that emerge when foundation models operate as autonomous agents with the ability to take actions, access external tools, and make irreversible decisions in dynamic, adversarial environments.

Distinct from prior workshops, which have largely focused on alignment or red-teaming of static foundation models, our workshop centers on the emerging safety, security, and trustworthiness challenges unique to agentic systems that act, adapt, and interact over time. By focusing on operational rather than purely theoretical safety, this workshop aims to bridge traditionally siloed research areas spanning AI safety, software security, and cyber-physical systems to define the foundations of secure and trustworthy autonomy in realistic, multi-agent, tool-augmented environments. To our knowledge, no existing venue provides a sustained, cross-domain focus on the technical and societal challenges of securing and governing agentic AI systems in real-world deployments.

The organizing committee is committed to supporting research related to safe, secure, and trustworthy agentic AI, and aims to continue this workshop series in the coming years as agentic AI becomes more capable, widespread, and deeply integrated into real-world systems and decision-making processes.

REFERENCES

[1] Yonadav Shavit et al. "Practices for governing agentic AI systems". In: *Research Paper, OpenAI* (2023).

[2] Yoshua Bengio et al. "Managing extreme AI risks amid rapid progress". In: *Science* 384.6698 (2024), pp. 842–845.

[3] Zhiheng Xi et al. *AgentGym-RL: Training LLM Agents for Long-Horizon Decision Making through Multi-Turn Reinforcement Learning*. 2025. arXiv: 2509.08755 [cs.LG]. URL: https://arxiv.org/abs/2509.08755.

[4] Shunyu Yao et al. *ReAct: Synergizing Reasoning and Acting in Language Models*. 2023. arXiv: 2210.03629 [cs.CL]. URL: https://arxiv.org/abs/2210.03629.

[5] Lutfi Eren Erdogan et al. *Plan-and-Act: Improving Planning of Agents for Long-Horizon Tasks*. 2025. arXiv: 2503.09572 [cs.CL]. URL: https://arxiv.org/abs/2503.09572.

[6] Yoshua Bengio et al. *International AI Safety Report*. 2025. arXiv: 2501.17805 [cs.CY]. URL: https://arxiv.org/abs/2501.17805.

[7] Beatrice Nolan. *AI-powered coding tool wiped out a software company's database in 'catastrophic failure'*. en. July 2025. URL: https://fortune.com/2025/07/23/ai-coding-tool-replit-wiped-database-called-it-a-catastrophic-failure/.

[8] Thor Olavsrud. *11 famous AI disasters*. en. Aug. 2025. URL: https://www.cio.com/article/190888/5-famous-analytics-and-ai-disasters.html.

[9] Yoshua Bengio et al. "Superintelligent agents pose catastrophic risks: Can scientist ai offer a safer path?" In: *arXiv preprint arXiv:2502.15657* (2025).

[10] Yoshua Bengio et al. *The Singapore Consensus on Global AI Safety Research Priorities*. 2025. arXiv: 2506.20702 [cs.AI]. URL: https://arxiv.org/abs/2506.20702.

[11] Lewis Hammond et al. "Multi-agent risks from advanced ai". In: *arXiv preprint arXiv:2502.14143* (2025).

[12] Shaokun Zhang et al. "Which Agent Causes Task Failures and When? On Automated Failure Attribution of LLM Multi-Agent Systems". In: *arXiv preprint arXiv:2505.00212* (2025).

[13] Edoardo Debenedetti et al. *AgentDojo: A Dynamic Environment to Evaluate Prompt Injection Attacks and Defenses for LLM Agents*. 2024. arXiv: 2406.13352 [cs.CR]. URL: https://arxiv.org/abs/2406.13352.

[14] Aengus Lynch et al. "Agentic Misalignment: How LLMs Could be an Insider Threat". In: *Anthropic Research* (2025). https://www.anthropic.com/research/agentic-misalignment.

[15] Yifei Wang, Dizhan Xue, Shengjie Zhang, and Shengsheng Qian. *BadAgent: Inserting and Activating Backdoor Attacks in LLM Agents*. 2024. arXiv: 2406.03007 [cs.CL]. URL: https://arxiv.org/abs/2406.03007.

[16] Alessandro Cui, Alessio Lo Duca, and Lorenzo Cavallaro. "The Dark Side of LLMs: Agent-based Attacks for Complete Computer Takeover". In: *arXiv preprint arXiv:2507.06850* (2024).

[17] Zhensu Li, Yu Wang, Chao Liu, and Yang Liu. "Inducing Vulnerable Code Generation in LLM Coding Assistants". In: *arXiv preprint arXiv:2504.15867* (2024).

[18] Shenao Yan et al. "An LLM-Assisted Easy-to-Trigger Backdoor Attack on Code Completion Models: Injecting Disguised Vulnerabilities against Strong Detection". In: *33rd USENIX Security Symposium (USENIX Security 24)*. 2024, pp. 1795–1812.

[19] Zhun Wang et al. *AgentVigil: Generic Black-Box Red-teaming for Indirect Prompt Injection against LLM Agents*. 2025. arXiv: 2505.05849 [cs.CR]. URL: https://arxiv.org/abs/2505.05849.

[20] Yejin Bang et al. "HalluLens: LLM Hallucination Benchmark". In: (2025). arXiv: 2504.17550 [cs.CL]. URL: https://arxiv.org/abs/2504.17550.

[21] Diego Gosmar and Deborah A Dahl. "Hallucination mitigation using agentic ai natural language-based frameworks". In: *arXiv preprint arXiv:2501.13946* (2025).

[22] Zhun Wang et al. *CyberGym: Evaluating AI Agents' Real-World Cybersecurity Capabilities at Scale*. 2025. arXiv: 2506.02548 [cs.CR]. URL: https://arxiv.org/abs/2506.02548.

[23] Vineeth Sai Narajala and Om Narayan. "Securing agentic ai: A comprehensive threat model and mitigation framework for generative ai agents". In: *arXiv preprint arXiv:2504.19956* (2025).

[24] Yifeng He, Ethan Wang, Yuyang Rong, Zifei Cheng, and Hao Chen. "Security of ai agents". In: *2025 IEEE/ACM International Workshop on Responsible AI Engineering (RAIE)*. IEEE. 2025, pp. 45–52.

[25] Mert Cemri et al. "Why Do Multi-Agent LLM Systems Fail?" In: *arXiv preprint arXiv:2503.13657* (2025).

[26] Ohav Barbi, Ori Yoran, and Mor Geva. *Preventing Rogue Agents Improves Multi-Agent Collaboration*. 2025. arXiv: 2502.05986 [cs.CL]. URL: https://arxiv.org/abs/2502.05986.

[27] Xiangyu Qi et al. "Fine-tuning Aligned Language Models Compromises Safety, Even When Users Do Not Intend To!" In: *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024. URL: https://openreview.net/forum?id=hTEGyKf0dZ.

[28] Simon Lermen, Charlie Rogers-Smith, and Jeffrey Ladish. *LoRA Fine-tuning Efficiently Undoes Safety Training in Llama 2-Chat 70B*. 2023. URL: https://arxiv.org/abs/2310.20624.

[29] Jan Betley et al. *Emergent Misalignment: Narrow finetuning can produce broadly misaligned LLMs*. 2025. arXiv: 2502.17424 [cs.CL]. URL: https://arxiv.org/abs/2502.17424.