AGENTIC AI IN THE WILD: FROM HALLUCINATIONS TO RELIABLE AUTONOMY

Website: https://hallucination-reliable-agentic-AI.github.io

1 SUMMARY AND GOALS

When we delegate tasks to AI agents—can we count on them to get it right? Agentic AI systems are increasingly stepping beyond static generation tasks into autonomous decision-making: scheduling meetings, booking travel, managing workflows, and assisting in scientific research. In these contexts, *reliability is not just important—it is essential*. Yet today's foundation models remain prone to a critical failure mode: hallucination, where outputs are factually incorrect, semantically implausible, or detached from reality (Maynez et al., 2020; Ji et al., 2023; Wachter et al., 2024; Huang et al., 2025). While hallucinations are concerning in any generative system, these challenges are ***amplified in agentic settings***, where models execute sequences of decisions without continuous human oversight.

The rise of agentic AI marks a critical inflection point: we are no longer simply interacting with models—we are increasingly entrusting them to act on our behalf. Unlike standard prompt-response interactions, agentic systems introduce temporal and operational autonomy, where reasoning and execution unfold independently of the user. Consider a scenario where a user instructs an AI agent to secure conference registration by a specific deadline. The agent must interpret requirements, research available options, make selections, and complete transactions. However, if the agent becomes unreliable during any part of this process—by fabricating confirmation details, misinterpreting the user's intent, or falsely reporting task completion—the failure may go unnoticed until much later, resulting in costly and irreversible consequences.

Topics and key questions. Despite the growing deployment of such systems, we lack a foundational understanding of what makes them reliable. This workshop will bring together researchers across machine learning, NLP, HCI, robotics, and AI safety to explore algorithmic advances, formal guarantees, empirical insights, and design frameworks for improving agent reliability. We welcome contributions on topics including hallucination detection, uncertainty estimation, safe planning under ambiguity, evaluation benchmarks for agents, and interaction protocols for trustworthy human-agent collaboration. Key questions include:

- How do hallucinations manifest in multi-step, autonomous agent workflows—and how can we detect them before harm is done?
- What forms of uncertainty quantification can forecast model failure in real-time agentic execution?
- How can we formalize "reliability" in agentic contexts where success is long-horizon and ambiguous?
- What strategies allow agents to defer, escalate, or communicate uncertainty effectively to human users?
- What types of benchmarks, platforms and metrics would be sufficient for documenting unreliability in agentic systems in the wild?

Societal impact. The rise of autonomous AI agents raises profound societal stakes. In domains such as healthcare, law, scientific research, and education, hallucinated outputs are not merely technical failures—they can lead to misinformation, loss of resources, legal liability, and erosion of public trust. As AI systems shift from tools we supervise to agents we rely on, reliability becomes an urgent public concern. Equipping agents with the ability to recognize their own uncertainty, avoid false confidence, and defer when necessary is essential for their safe deployment in high-stakes environments. This workshop addresses not just a research gap, but a societal responsibility: ensuring that future agentic AI systems are not only capable, but cautious, transparent, and trustworthy.

2 INVITED SPEAKERS

The invited speakers were selected based on their relevance and prior contributions to the field. Each speaker has authored high-impact publications in machine learning.

Dawn Song (UC Berkeley) [***confirmed***] is a Professor at the University of California, Berkeley. She is a leading expert in AI, security, and safety. Dawn has received numerous accolades, including the Alfred P. Sloan Research Fellowship, ACM Fellow, and a MacArthur Fellow. She has recently turned her focus to the reliability and safety of agentic AI, including foundational work on verifiable and controllable AI agents, which aligns well with the theme of the workshop.

Mohan Kankanhalli (National University of Singapore) [***confirmed***] is the Director of NUS AI Institute, Provost's Chair Professor and is also the Deputy Executive Chairman of AI Singapore (Singapore's national AI program). Mohan leads major initiatives on human-centric and reliable AI, including deploying agentic systems in real-world environments.

Hamed Hassani (University of Pennsylvania) [***confirmed***] is an associate professor at the University of Pennsylvania, and visiting faculty researcher at Google Research. His works span across LLM safety and foundations of ML and information theory, and therefore can offer unique perspectives on uncertainty quantification for LLM agents.

Florian Buettner (Goethe-University Frankfurt, and German Cancer Research Center) [***confirmed***] is a professor interested in increasing the precision of ML tools for oncology. Florian has worked on uncertainty estimation to make models more reliable.

James Zou (Stanford University) [***confirmed***] works on reliable machine learning with foundation models and is particularly interested for their applications to the health domain.

Stefano Soatto (UCLA and Vice president of AWS) [***confirmed***] works on representation and optimal control, including the control of AI agents and their reliability.

Besides the confirmed speakers, there is a list of tentative speakers that we can invite in case of cancellations (our original schedule posted below is complete):

- Csaba Szepesvári (University of Alberta and team lead at Deepmind, sequential decision making, uncertainty).
- Graham Neubig (Carnegie Mellon University, question answering, code generation, hallucination of LLMs).
- Sandra Wachter (University of Oxford, factuality of LLMs).
- Yu Su (Ohio State University, NLP and language agents)

3 SCHEDULE, CALL FOR PAPERS AND REVIEWING PROCESS

The workshop will span a **full day** and include a mix of invited talks, spotlight presentations of selected accepted papers, poster sessions for all accepted submissions, and panel discussion. **Three** poster sessions will be organized, following the positive feedback from our ICLR workshop (Chrysos et al., 2025) where we experimented with increased number of poster sessions. This led to an increased interaction and organic discussions between the interested participants. The tentative schedule is outlined below.

Call for papers and reviewing: We invite original research (theoretical, empirical) submissions on **trustworthy agentic AI and hallucinations**. The process will be managed on *OpenReview* with a **double-blind** policy. The timeline (Table 1) follows the official ICLR schedule. Each paper will receive $\geq$**3 expert reviews** plus a meta-review, evaluated on rigor, novelty, impact, reproducibility and relevance to our program on agentic AI and hallucinations. To ensure a high-quality and fair process, we will enforce a strict conflict-of-interest policy, cap reviewer workloads. We will select a balanced program across methods, domains, and modalities. We will also conduct targeted outreach (open call, mailing lists, affinity groups), and reserve slots for replications/negative results. We also have initiatives to foster an inclusive community, including a **peer mentorship program** for new reviewers and awards for **best paper and best artifact**. To maintain neutrality, the organizers will not give invited talks.

start	duration	event	theme
9:00	0:10	opening remarks	
9:10	0:35	invited talk: Florian Buettner	Hallucination in agentic systems
9:45	0:35	invited talk: Hamed Hassani	
10:20	1:00	poster session I	
11:20	0:35	invited talk: Dawn Song	
11:55	1:00	lunch	
12:55	1:00	poster session II	Towards reliable AI agents
13:55	0:30	invited talk: Stefano Soatto	
14:25	0:35	invited talk: James Zou	
15:00	0:30	invited talk: Mohan Kankanhalli	
15:30	1:00	poster session III	
16:30	0:10	Best paper award	
16:40	1:00	panel	
17:40	0:10	closing remarks	

Similarly, for conflicts of interest, e.g., paper from own group or PhD group or the same organization, the organizers will recuse themselves from the decision. Workshop reviewers will not give invited talks to avoid conflict of interest. We will **establish a tiny track of papers** - as workshop organizers we did this in our ICLR'25 workshop as well.

Table 1: Timeline for contributed work submissions

December 5th (2025)	Portal opens for submissions
December 15th (2025)	Mentoring sessions
January 30th (2026)	Submission deadline
February 20th (2026)	Reviewing deadline
February 26th (2026)	Notification date
April 20th (2026)	Camera-ready deadline

4 Audience, Access and Interactions

Audience: The topic of agentic AI and hallucinations in foundation models has received increasing attention, with numerous papers accepted at ICLR, ICML, and NeurIPS over the past two years. Hallucination has also attracted the attention in the popular media (Weisse & Metz, 2023; Weiser, 2023; Kaye, 2023; Verma & Oremus, 2023; Kelsey-Sugg & Carrick, 2024; Legg & McNamara, 2024; Brittain, 2025; Rahman-Jones, 2025; Weaver, 2025), as explicitly identified by judges (IANS, 2023) or in the medical domain (Hughes, 2024), with the problem becoming worse with stronger models (Metz & Weisse, 2025). Consequently, we anticipate substantial interest from the community in a dedicated workshop on this subject. We expect participation from both early career researchers and students, as well as established researchers seeking fresh insights into this emerging field. Therefore, we estimate approximately 300 participants for the workshop.

Website and Access: Our website https://hallucination-reliable-agentic-AI.github.io will release all accepted papers at least one week prior to the workshop, allowing participants to review the papers in advance. Following ICLR guidelines, our primary focus will be on in-person attendance. However, to accommodate extenuating circumstances such as visa issues or other rare exceptions, we will allow online participation for people that have a valid reason.

Cross-community interaction: We will intentionally engineer mixing between machine learning, HCI, planning/agents, LLM and safety communities: (i) paper–discussant pairing that assigns each accepted paper a discussant from a different field, with a short cross-disciplinary response after the talk/poster; (ii) guided poster walks where moderators lead mixed-background groups through themed clusters (tool use, evaluation, human studies, theory), using prompt cards to surface divergent assumptions; (iii) cross-review: at least one reviewer per paper comes from an adjacent area; and (iv) ongoing Slack/Discord with channels by theme (ML and AI, theory, human studies, planning, eval) to coordinate post-workshop projects and shared task proposals.

5 Related Workshops

Beyond ICLR, there have been no workshops on reliable agentic AI at ICML or NeurIPS. The most closely related event was the ICLR workshop on uncertainty quantification in foundation models (Chrysos et al., 2025). That workshop shared some organizers as the current proposal, but had a different focus. Many of the organizers of the previous workshop focus on uncertainty quantification, which is different than the present proposal. The theme of the workshop is also different with the current proposal focusing on the significant and emerging topic of agentic AI that was not explored or

mentioned at all in ICLR'25. The ICLR'25 workshop accepted 46 papers and experienced significant attendance, with invited talks drawing audiences beyond the room's 200-person capacity. It briefly mentioned hallucinations in text models (in the panel questions) and did not cover AI agents. By contrast, our workshop puts higher emphasis on agentic settings, where the agentic AI systems can perceive, plan, and act with tools, memory, and multi-step reasoning. This moves uncertainty beyond next-token probability to the whole decision pipeline—reasoning traces, tool calls, retrieved evidence, and downstream actions, which is both fundamentally significant and practically important to the broader audience. As such, we do not consider that there is a significant overlap with the proposed content.

Other related workshops include workshops on LLM agents for science (Chen & et al., 2024; Koutra & et al., 2025), reinforcement learning agents (Jiang & et al., 2022; de Witt & et al., 2023), out-of-distribution learning (Deshmukh & et al., 2023), or specific types of generative models (Thanapalasingam & et al., 2023). The closest workshops are those on trustworthy ML (Papernot & et al., 2020; Beirami & et al., 2021; Xiao & et al., 2022; Cheng & et al., 2023; Bansal & et al., 2024). None of them focus on hallucinations and they mostly refer to adversarial robustness, which is a more narrow topic than the questions we will face with agentic systems. Therefore, we believe the proposed workshop does not have a significant overlap with the past workshops.

6 Diversity of Team and Topics

Organizing Committee Diversity: The organizing committee reflects a broad range of backgrounds, with deliberate attention to diversity in gender, career stage, institutional affiliation, and geographic representation. The core team consists of 1 applied scientist, 2 assistant professors, 1 associate professor and 1 research professor. Two additional volunteers will assist with workshop logistics. The committee includes 5 women (representing 70% of the organizing committee), underscoring a commitment to gender representation. Members are affiliated with institutions in the United States, and Singapore, spanning academia, research institutes and industry. This composition is intended to ensure a balanced and inclusive environment.

Invited Speaker Diversity The invited speakers represent a mix of senior and early-career researchers. Their contributions cover both theoretical and empirical research with practical applications. A common focus across all speakers is the development of reliable and trustworthy AI systems for deployment in real-world contexts, while the invited speakers are affiliated with institutes in both US and Europe.

Theme Diversity The workshop is structured to promote interdisciplinary engagement, drawing on expertise from applied mathematics, statistics, computer science, HCI, and machine learning. Emphasis will be placed on the practical integration of these disciplines, with the goal of fostering new collaborations and advancing understanding across traditional boundaries.

7 Organizers

Below, we provide further details on the core team and the program committee of the workshop.

7.1 Core team

The organizing team has made significant contributions to various areas of out-of-distribution prediction, trustworthy models, uncertainty quantification and generative models. In addition to their scholarly work, they bring extensive organizational experience, having led more than a **fiteen workshops and tutorials** over the past decade. These efforts have successfully engaged a wide range of sub-communities within the broader machine learning. Several members of the team have also held prominent leadership roles, including **serving as Program Chairs for ICML**.

Grigorios Chrysos (UW-Madison) is an Assistant Professor at the University of Wisconsin-Madison. Before that, Grigorios was a postdoctoral fellow at EPFL following the completion of his PhD at Imperial College London. Previously, he graduated from the National Technical University of Athens with a Diploma/MEng in Electrical and Computer Engineering. Grigorios has co-organized workshops in top conferences (ICCV'15, CVPR'17, ICCV'17, NeurIPS'24, AAAI'25, ICLR'25). The most recent NeurIPS workshop ('*Fine-Tuning in Modern Machine Learning: Principles and Scalability*) is complementary to this workshop and focuses on different perspectives of foundation

models. Grigorios also organized tutorials on tensors and architecture design (CVPR'22, AAAI'23, DSAA'24, NeurIPS'25, Dagstuhl'26) and deep learning theory (CVPR'23, ISIT'24). His research interests lie in trustworthy machine learning, publishing many results in top-tier conferences (CVPR, NeurIPS, ICLR, ICML). Grigorios serves as an Associate Editor for TMLR and an Area Chair for top-tier ML conferences (ICLR, ICML, NeurIPS). [Google Scholar][Email]

Sharon Li (UW-Madison) is an Associate Professor in the Department of Computer Sciences at the University of Wisconsin-Madison. She received a Ph.D. from Cornell University in 2017, advised by John E. Hopcroft. Subsequently, she was a postdoctoral fellow in the Computer Science department at Stanford University. Her research focuses on the algorithmic and theoretical foundations of reliable machine learning in the open world, as well as developing responsible foundation models, including large language models and vision-language models. Sharon has served as the founding organizer and Program Chair for the *ICML Workshop on Uncertainty and Robustness in Deep Learning* (2019 and 2020), co-organized multiple other workshops including the *ICML Workshop on Distribution-free Uncertainty Quantification* in 2021 and 2022, *NeurIPS'22 Workshop on Robustness in Sequence Modeling*, and *ICCV'23 Tutorial on Reliability of Deep Learning for Real-World Deployment*, and *ICLR'25 Workshop on Quantify Uncertainty and Hallucination in Foundation Models: The Next Frontier in Reliable AI*, She has served as Area Chair and Senior Program Committee for top-tier ML conferences including ICLR, ICML and NeurIPS between 2020 and 2024. She is serving as the Program Chair of ICML 2026. [Google Scholar][Email]

Etsuko Ishii (Amazon) is an Applied Scientist at Amazon Web Services, focusing on developing and advancing Agentic AI applications. Her research expertise lies in contextual understanding of dialogues and building agentic systems that can effectively interact with users. Before joining AWS, she completed her Ph.D. in Electronic & Computer Engineering at the Hong Kong University of Science and Technology in 2024. She serves as Area Chair for ACL Rolling Review. [Google Scholar][Email]

Sean Du (Nanyang Technological University - Singapore) is an Assistant Professor at College of Computing and Data Science (CCDS), Nanyang Technological University, Singapore. He obtained his Ph.D. in Computer Sciences at UW-Madison. His research interest is in reliable machine learning and the applications to foundation models and AI safety. His first-author papers have been recognized with multiple oral and spotlight presentations at NeurIPS and CVPR. He is a recipient of the Jane Street Graduate Research Fellowship, and Rising Stars in Data Science award. Sean has co-organized the *IJCNN'25 Special Session on Responsible Foundation Models in the Wild*. [Google Scholar][Email]

Katia Sycara (CMU) is a Research Professor at Carnegie Mellon University. She is a Fellow of IEEE and AAAI, with over 300 publications in multi-agent systems, semantic web, and human-agent interaction. Her work has been funded by DARPA, NASA, NSF, and others, including the development of the RETSINA multi-agent infrastructure. She has held leadership roles in major conferences and contributed to web standards through W3C and OASIS. Sycara has received multiple honors, including the ACM/SIGART Agents Research Award and an honorary doctorate. Prof. Sycara is a founding member and member of the Board of Directors of the International Foundation of Multiagent Systems (IFMAS). She is a founding Editor-in-Chief of the journal "Autonomous Agents and Multiagent Systems"; an Editor-in-Chief of the Springer Series on Agents; on the Editorial Board of the Kluwer book series on "Multiagent Systems, Artificial Societies and Simulated Organizations". [Google Scholar][Email]

7.2 Program Committee

Volunteers: Two student (one PhD student and one undergraduate student) will help the organizers hosting the workshop: Andrea Tseng (University of Wisconsin-Madison) and Yiheng Zhang (University of Wisconsin-Madison).

In addition, the following people have agreed to serve in the program committee:

Elias Rocamora (EPFL), Blerina Gkotse (University of Wisconsin-Madison), Justin Deschenaux (EPFL), Andrea Tseng (University of Wisconsin-Madison), Thomas Pethick (EPFL), Stratis Skoulakis (Aarhus university), Dimitris Halatsis (Imperial College London), Muhammad Ashiq (University of Wisconsin-Madison), Zhiyuan Wu (University of Oslo), Aggelina Chatziagapi (Stonybrook University), Jiankang Deng (Imperial College London), Seongheon Park (University of Wisconsin-

Madison), Changdae Oh (University of Wisconsin-Madison), Froilan Choi (University of Wisconsin-Madison), Shaokun Zhang (Penn State University), Jiachen (Tianhao) Wang (Princeton University), Yiran Wu (Penn State University), Wenyue Hua (Microsoft Research), Zhouxing Shi (UC Riverside).

Depending on the number of submissions received, it may become necessary to expand the program committee. Given the track record in the past year, the core team's extensive experience in organizing scientific events will be instrumental in mobilizing their broad professional network to support the review process. **Concretely, as organizers we have organized before workshops scaling to hundreds of submissions and we recruited additional reviewers from our networks.**

REFERENCES

Bansal and et al. Workshop on 'reliable and responsible foundation models'. In *International Conference on Learning Representations (ICLR) Workshops*, 2024. [URL: https://iclr.cc/virtual/2024/workshop/20588].

Beirami and et al. Workshop on 'responsible ai (rai)'. In *International Conference on Learning Representations (ICLR) Workshops*, 2021. [URL: https://iclr.cc/virtual/2021/workshop/2132].

Blake Brittain. Anthropic's lawyers take blame for ai 'hallucination' in music publishers' lawsuit, 2025. [Data: https://www.reuters.com/legal/legalindustry/anthropics-lawyers-take-blame-ai-hallucination-music-publishers-lawsuit-2025-05-15/. Status: Online; accessed 9-October-2025].

Chen and et al. Workshop on 'large language models for agents'. In *International Conference on Learning Representations (ICLR) Workshops*, 2024. [URL: https://iclr.cc/virtual/2024/workshop/20575].

Cheng and et al. Workshop on 'trustworthy and reliable large-scale machine learning models'. In *International Conference on Learning Representations (ICLR) Workshops*, 2023. [URL: https://iclr.cc/virtual/2023/workshop/12827].

Grigorios Chrysos, Sharon Li, Anastasios Angelopoulos, Stephen Bates, Barbara Plank, and Emtiyaz Khan. Workshop on 'quantify uncertainty and hallucination in foundation models: The next frontier in reliable ai'. In *International Conference on Learning Representations (ICLR) Workshops*, 2025. [URL: https://uncertainty-foundation-models.github.io].

Schroeder de Witt and et al. Workshop on 'ai for agent-based modelling (ai4abm)'. In *International Conference on Learning Representations (ICLR) Workshops*, 2023. [URL: https://iclr.cc/virtual/2023/workshop/12840].

Deshmukh and et al. Workshop on 'what do we need for successful domain generalization?'. In *International Conference on Learning Representations (ICLR) Workshops*, 2023. [URL: https://iclr.cc/virtual/2023/workshop/12824].

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, et al. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 43(2):1–55, 2025.

Alex Hughes. Your gp could be using chatgpt to diagnose you, a study finds, 2024. [Data: https://www.sciencefocus.com/news/gps-use-chatgpt-study-finds. Status: Online; accessed 9-October-2025].

IANS. Us judge orders lawyers not to use chatgpt-drafted content in court, 2023. [Data: https://www.business-standard.com/world-news/us-judge-orders-lawyers-not-to-use-chatgpt-drafted-content-in-court-123053100238_1.html. Status: Online; accessed 9-October-2025].

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. Survey of hallucination in natural language generation. *ACM computing surveys*, 55(12): 1–38, 2023.

Jiang and et al. Workshop on 'agent learning in open-endedness'. In *International Conference on Learning Representations (ICLR) Workshops*, 2022. [URL: https://iclr.cc/virtual/2022/workshop/4547].

Byron Kaye. Australian mayor readies world's first defamation lawsuit over chatgpt content, 2023. [Data: https://www.reuters.com/technology/australian-mayor-readies-worlds-first-defamation-lawsuit-over-chatgpt-content-2023-04-05/. Status: Online; accessed 9-October-2025].

Anna Kelsey-Sugg and Damien Carrick. Ai hallucinations caused artificial intelligence to falsely describe these people as criminals, 2024. [Data: https://www.abc.net.au/news/2024-11-04/ai-artificial-intelligence-hallucinations-defamation-chatgpt/104518612. Status: Online; accessed 9-October-2025].

Koutra and et al. Workshop on 'towards agentic ai for science: Hypothesis generation, comprehension, quantification, and validation'. In *International Conference on Learning Representations (ICLR) Workshops*, 2025. [URL: https://iclr.cc/virtual/2025/workshop/23991].

Michael Legg and Vicki McNamara. Ai is creating fake legal cases – and making its way into real courtrooms with disastrous results, 2024. [Data: https://theconversation.com/ai-is-creating-fake-legal-cases-and-making-its-way-into-real-courtrooms-with-disastrous-res Status: Online; accessed 9-October-2025].

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. On faithfulness and factuality in abstractive summarization. *arXiv preprint arXiv:2005.00661*, 2020.

Cade Metz and Karen Weisse. A.i. hallucinations are getting worse as models get smarter, 2025. [Data: https://www.nytimes.com/2025/05/05/technology/ai-hallucinations-chatgpt-google.html. Status: Online; accessed 9-October-2025].

Papernot and et al. Workshop on 'towards trustworthy ml: Rethinking security and privacy for ml'. In *International Conference on Learning Representations (ICLR) Workshops*, 2020. [URL: https://iclr.cc/virtual_2020/workshops_6.html].

Imran Rahman-Jones. Man files complaint after chatgpt said he killed his children, 2025. [Data: https://www.bbc.com/news/articles/c0kgydkr516o. Status: Online; accessed 9-October-2025].

Thanapalasingam and et al. Workshop on 'neurosymbolic generative models (nesy-gems)'. In *International Conference on Learning Representations (ICLR) Workshops*, 2023. [URL: https://iclr.cc/virtual/2023/workshop/12830].

Pranshu Verma and Will Oremus. Chatgpt sometimes makes up facts. for one law prof, it went too far, 2023. [Data: https://www.washingtonpost.com/technology/2023/04/05/chatgpt-lies/. Status: Online; accessed 9-October-2025].

Sandra Wachter, Brent Mittelstadt, and Chris Russell. Do large language models have a legal duty to tell the truth? *Royal Society Open Science*, 11(8):240197, 2024.

Matthew Weaver. Ai chatbots distort and mislead when asked about current affairs, bbc finds, 2025. [Data: https://www.theguardian.com/technology/2025/feb/11/ai-chatbots-distort-and-mislead-when-asked-about-current-affairs-bbc-finds. Status: Online; accessed 9-October-2025].

Benjamin Weiser. Here's what happens when your lawyer uses chatgpt, 2023. [Data: https://www.nytimes.com/2023/05/27/nyregion/avianca-airline-lawsuit-chatgpt.html. Status: Online; accessed 9-October-2025].

Karen Weisse and Cade Metz. When a.i. chatbots hallucinate, 2023. [Data: https://www.nytimes.com/2023/05/01/business/ai-chatbots-hallucination.html. Status: Online; accessed 9-October-2025].

Xiao and et al. Workshop on 'socially responsible machine learning'. In *International Conference on Learning Representations (ICLR) Workshops*, 2022. [URL: https://iclr.cc/virtual/2022/workshop/4558].