Knowledge Control for Responsible Generative AI: Bridging Academia, Industry, and Society

Zheyuan Liu, Yixin Wan, Kai-Wei Chang, Meng Jiang, Jieyu Zhao, Nouha Dziri, Yuning Mao, Jia-Chen Gu, Jindong Gu

https://franciscoliu.github.io/knowledge-control-tutorial-2026/

Controlling the knowledge and behavior of generative AI systems, including large language models (LLMs), multimodal LLMs (MLLMs), and text-to-image (T2I) models, has become critical as they are increasingly used in safety-sensitive and socially impactful applications. These models often encode unintended, biased, or private content, leading to harmful or unethical outputs. Post-training knowledge control has thus emerged as a practical framework for selectively modifying or removing model behaviors without full retraining, offering scalable and interpretable interventions for improving safety, privacy, and fairness. This tutorial introduces the foundations of post-training knowledge control and showcases recent frontier methods, bridging research insights with real-world practices from both academia and industry. We cover: (i) key motivations and failure modes, such as harmful generation and stereotype reinforcement; (ii) core methods such as machine unlearning, knowledge editing, and inference-time interventions for targeted behavior adjustment; and (iii) evaluation protocols for balancing forgetting, retention, and fairness. Case studies will span text and vision–language generation, including privacy preservation, bias mitigation, and factual correction.

Zheyuan Liu
email: zliu29@nd.edu
website: https://franciscoliu.github.io/
Zheyuan Liu is a Ph.D. candidate in Computer Science and Engineering at the University of Notre Dame, advised by Prof. Meng Jiang. His research focuses on trustworthy generative AI, including LLM/MLLM safety, AI privacy, fairness, agentic safety, and multimodal models. He received his B.S. degrees in Computer Science and Applied Mathematics from Brandeis University.

Yixin Wan
email: elaine1wan@g.ucla.edu
website: https://scholar.google.com/citations?user=hZPIICQAAAAJ&hl=en
Yixin Wan is a Ph.D. Candidate in Computer Science at the University of California, Los Angeles, advised by Prof. Kai-Wei Chang. Her research focuses on trustworthy NLP and multimodal generative models, including fairness, safety, text-to-image generation, vision-language models, and machine unlearning. She received her undergraduate training in mathematics at UCLA.

Kai-Wei Chang
email: kwchang@cs.ucla.edu
website: https://web.cs.ucla.edu/~kwchang/
Kai-Wei Chang is a Professor of Computer Science at the University of California, Los Angeles. He co-directs the UCLA DataX AI Technology Center and leads research on trustworthy AI, multimodal foundation models, reasoning, controllable generation, and human-centered AI systems. His work spans fairness, robustness, unlearning, safety, and reliable NLP and multimodal systems.

Meng Jiang
email: mjiang2@nd.edu

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 5: Tutorial Abstracts), pages 9–10

website: http://www.meng-jiang.com/
Meng Jiang is the Frank M. Freimann Collegiate Professor of Computer Science and Engineering at the University of Notre Dame. His research interests include data mining, machine learning, natural language processing, artificial intelligence, and trustworthy generative AI. He has worked on topics such as computational behavior modeling, personalized language models, and machine unlearning.

Jieyu Zhao
email: jieyuz@usc.edu
website: https://jieyuz.net/
Jieyu Zhao is a Gabilan Assistant Professor in the Thomas Lord Department of Computer Science at the University of Southern California. She leads the LIME Lab, and her research focuses on trustworthy language models and human-centered AI. Before joining USC, she was an NSF Computing Innovation Fellow at the University of Maryland, College Park, and she received her Ph.D. in Computer Science from UCLA.

Nouha Dziri
email: nouha.dziri@gmail.com
website: https://nouhadziri.github.io/
Nouha Dziri is a Senior Research Scientist at Cohere Labs, after previously working as a research scientist and postdoctoral researcher at the Allen Institute for AI. Her research focuses on NLP and machine learning, with an emphasis on large language models, reasoning, safety, security, and post-training methods. Her work includes studies of hallucination, compositional reasoning, red-teaming, and safety-oriented post-training for open language models.

Yuning Mao
email: yuningm@meta.com
website: https://morningmoni.github.io/
Yuning Mao is a Research Scientist at Meta Superintelligence Labs. His work focuses on large language models, post-training, code generation, and helping users acquire information and knowledge more effectively. He received his Ph.D. in Computer Science from the University of Illinois Urbana-Champaign and has contributed to Meta's Llama series and related post-training efforts.

Jia-Chen Gu
email: gujc@ucla.edu
website: https://jasonforjoy.github.io/
Jia-Chen Gu is a Postdoctoral Researcher in the Department of Computer Science at the University of California, Los Angeles. He is hosted by Prof. Nanyun Peng and Prof. Kai-Wei Chang, and he received his Ph.D. from the University of Science and Technology of China in 2022. His research focuses on machine learning for natural language processing, especially retrieval-augmented language models, model editing, representation learning, and robust, efficient, and trustworthy AI.

Jindong Gu
email: jindong.gu@outlook.com
website: https://jindonggu.github.io/
Jindong Gu is a Senior Research Scientist at Google and a Senior Research Fellow at the University of Oxford. His research focuses on responsible AI, including the reliability, robustness, privacy, interpretability, and safety of AI systems. He received his Ph.D. from the University of Munich in 2022 and has worked on the safety of foundation models and the robustness of AI agents.