FROM HUMAN COGNITION TO AI REASONING: MODELS, METHODS, AND APPLICATIONS

Julie A. Shah[1] **Sarath Sreedharan**[2] **Silvia Tulli**[3] **Pulkit Verma**[1,4]
[1]Massachusetts Institute of Technology, USA
[2]Colorado State University, USA
[3]Sorbonne University, France
[4]Indian Institute of of Technology Madras, India

https://bit.ly/hcair26

1 INTRODUCTION

The objective of this workshop is to bridge the gap between human cognitive science and artificial intelligence by bringing together researchers working on computational models of human cognition, neurosymbolic AI, human-AI interaction, and cognitively-inspired machine learning. Recent advances in AI have demonstrated remarkable capabilities, yet these systems often lack the interpretability, causal reasoning, and generalization abilities that characterize human intelligence. Meanwhile, cognitive science has made significant progress in understanding human reasoning, learning, and decision-making processes. We believe that incorporating insights from human cognition into AI systems can lead to more robust, interpretable, and human-aligned artificial intelligence. This workshop aims to facilitate cross-pollination of ideas between cognitive scientists, neuroscientists, and AI researchers to develop the next generation of AI systems that can reason more like humans while maintaining computational efficiency.

The workshop will explore how explicit models of human knowledge, cognitive capabilities, and mental states can be integrated into AI reasoning processes. We will examine approaches that combine neural and symbolic methods inspired by human cognition, incorporate human causal reasoning patterns, and leverage human teaching signals to create more interpretable and aligned AI systems.

ICLR presents an ideal, inclusive venue for dialogue and technical interaction among researchers spanning cognitive science, machine learning, AI planning, human-robot interaction, and neurosymbolic AI communities.

1.1 WORKSHOP TOPICS

The workshop will focus on research related to all aspects of human cognition and AI reasoning. This topic features technical problems that are of interest across multiple fields including cognitive science, machine learning, AI planning, human-robot interaction, and neurosymbolic AI. We welcome submissions that address **formal** as well as **empirical** issues on topics such as:

- Explicit modeling of human knowledge and cognitive capabilities
- Introducing explicit human models into the reasoning process
- How AI can model and reason about human mental states and intentions
- Combining neural and symbolic methods inspired by human cognition
- Incorporating human causal reasoning patterns into AI systems
- Using human cognitive models to make AI systems more interpretable
- Incorporating human teaching and correction into learning processes
- Structured approaches to human-inspired AI reasoning
- Probabilistic approaches to human-like reasoning

Attendance Based on the interdisciplinary nature of this workshop and the growing interest in cognitively-inspired AI, we expect **80–120 attendees** at ICLR.

2 Workshop Format

We propose to organize a one-day workshop featuring presentations of **contributed papers**, a **poster session**, **invited plenary talks**, and an inclusivity focused **focused discussion groups** session. focused discussion groups are designed as safe spaces for junior researchers and researchers from under-represented communities to voice their ideas, opinions, and impressions in smaller supportive groups with senior members from the community. The design of this session is discussed further under "Diversity Statement."

2.1 Submission Logistics

We invite contributions in multiple formats to encourage diverse participation. Full research papers should adhere to the standard ICLR formatting guidelines with a length between 4 and 9 pages. Authors are also welcome to submit early-stage research, including preliminary findings accompanied by demonstrations. Additionally, we will support a **blue sky ideas track** for visionary position papers that propose bold, speculative directions for future research at the intersection of human cognition and AI reasoning. These blue sky submissions may present novel paradigms, thought experiments, or unconventional approaches without requiring extensive empirical validation.

This workshop operates as a non-archival venue, meaning accepted contributions will not appear in formal conference proceedings. Notification of acceptance will be communicated to authors by 1 March 2026, after which point all accepted manuscripts will be publicly accessible through our workshop webpage. To promote original scholarship, we require that submitted work has not previously appeared at any machine learning venue, including but not limited to the main ICLR conference track.

A dedicated poster session will provide an interactive forum where contributors can engage in extended discussions with workshop participants. This format facilitates meaningful exchanges that can refine ongoing research and expose attendees to innovative methodologies and perspectives.

2.2 Post-workshop Content

We plan to post the workshop content online for wider dissemination. In addition to adding the papers on the workshop website, we will put the videos and slides for the contributed and invited talks, and posters for all the accepted papers online as well.

3 Invited Speakers

Invited speakers who have already been confirmed and have indicated a strong desire to attend and give invited talks:

Rachid Alami, CNRS Senior Scientist Emeritus, LAAS-CNRS, France.

He received an engineering diploma in computer science in 1978 from ENSEEIHT, a Ph.D in Robotics in 1983 from Institut National Polytechnique and an Habilitation HDR in 1996 from Paul Sabatier University. His main research contributions fall in the fields of Robot Decisional and Control Architectures, Task and motion planning, multi-robot cooperation, and human-robot interaction. Rachid Alami is currently the head of the Robotics and InteractionS group at LAAS. He has been offered in 2019 the Academic Chair on Cognitive and Interactive Robotics at the Artificial and Natural Intelligence Toulouse Institute (ANITI).

Kimberly Lauren Stachenfeld, Senior Research Scientist, Google DeepMind, USA and Adjunct Assistant Professor, Columbia University, USA.

She completed her PhD in Neuroscience at Princeton University where she was advised by Matthew Botvinick. Before that, she received my Bachelors from Tufts University in Mathematics (BA) and Chemical & Biological engineering (BS). Her research covers topics in

Neuroscience and AI. On the Neuroscience side, she studies how brains build models of the world that support memory and prediction. On the Machine Learning side, she works on implementing these cognitive functions in deep learning models. Her work has been featured in The Atlantic, Quanta Magazine, Nautilus, and MIT Technology Review. In 2019, she was named one of MIT Tech Review's Innovators under 35.

Joshua B Tenenbaum, Professor of Computational Cognitive Science, MIT, USA.

He received his PhD from MIT in 1999, and was an Assistant Professor at Stanford University from 1999 to 2002 before returning to MIT. His research centers on perception, learning, and common-sense reasoning in humans and machines, with the twin goals of better understanding human intelligence in computational terms and building more human-like intelligence in machines. His papers have received awards at the Cognitive Science (CogSci), Computer Vision and Pattern Recognition (CVPR), Neural Information Processing Systems (NIPS), and Uncertainty in Artificial Intelligence (UAI) conferences, the International Conference on Learning and Development (ICDL) and the International Joint Conference on Artificial Intelligence (IJCAI). He has given invited keynote talks at all of the major machine learning and artificial conferences.

Elmira Yadollahi, Assistant Professor, Lancaster University, UK.

She did her PhD as part of a joint doctoral initiative between École Polytechnique Fédérale de Lausanne (EPFL) in Switzerland and Instituto Superior Técnico in Portugal. Her doctoral dissertation was advised by Prof. Ana Paiva, the head of the GAIPS, and Prof. Pierre Dillenbourg, who directs the CHILI lab. Before that, she was a Postdoctoral Fellow at the Division of Robotics, Perception, and Learning (RPL) at KTH Royal Institute of Technology in Sweden since October 2021. Her research is focused on human-robot interaction, explainability in robotics, interaction design, and child-robot interaction. She is an associate editor of the International Journal of Child-Computer Interaction (IJCCI) and has served in several conferences as PC/SPC member.

4 Tentative Schedule

This schedule is subject to change based on the final group of participants and the number of contributed papers.

08:55 am - 09:00 am	Opening Remarks
09:00 am - 09:35 am	Invited Talk 1
09:35 am - 10:20 am	Contributed Papers
10:20 am - 10:30 am	Discussion
10:30 am - 11:00 am	Coffee Break
11:00 am - 11:35 am	Invited Talk 2
11:35 am - 12:20 pm	Contributed Papers
12:20 pm - 12:30 pm	Discussion
12:30 pm - 02:00 pm	Lunch Break
02:00 pm - 02:35 pm	Invited Talk 3
02:35 pm - 03:10 pm	Invited Talk 4
03:10 pm - 03:30 pm	Poster Session
03:30 pm - 04:00 pm	Coffee Break with Posters
04:00 pm - 04:10 pm	Poster Session
04:10 pm - 05:00 pm	Contributed Talks
05:00 pm - 05:30 pm	Panel Discussion

5 Diversity Statement

Our workshop is committed to bringing together researchers from diverse backgrounds and creating an inclusive environment for discussion.

The organizers will invite speakers, PC members, and encourage participation from underrepresented groups. We also plan to reach out to relevant researchers in underrepresented groups directly for active participation in the workshop through groups like Women in Machine Learning (WiML), Black in AI, LatinX in AI (LXAI), etc. To foster open communication and inclusiveness, we will adopt the same code of conduct and code of ethics guidelines as that of ICLR. In addition, we will organize inclusive focused discussion groups as follows.

*Inclusive engagement through **focused discussion groups**.* It is well-established that lower engagement is one of the major factors leading to poor inclusivity and that it can compel individuals from under-represented groups to withdraw from the broader community. We will organize a 60-minute session for *focused discussion groups* to increase engagement and help researchers from under-represented communities. This segment is designed to create a safe space where researchers from all backgrounds and all levels of experience will be able to express their points of view and build personal connections with the broader community.

Toward the end of the workshop, we will organize a general discussion using as starting points, the challenges and opportunities identified during the focused discussion groups. Participants will be welcome to contribute during the general discussion with any topic of their interest as long as it is aligned with the topics of the workshop.

6 Organizers

6.1 Organization Team

The organizing committee has been selected to build synergies across different research perspectives and communities addressing human cognition and AI reasoning. All committee members have published peer-reviewed work on the main topics of this workshop.

Julie A. Shah (julie_a_shah@csail.mit.edu)

Julie A. Shah is the H.N. Slater Professor and Head of Aeronautics and Astronautics, faculty director of MIT's Industrial Performance Center, and director of the Interactive Robotics Group, which aims to imagine the future of work by designing collaborative robot teammates that enhance human capability. She is expanding the use of human cognitive models for artificial intelligence and has translated her work to manufacturing assembly lines, healthcare applications, transportation and defense. Before joining the faculty, she worked at Boeing Research and Technology on robotics applications for aerospace manufacturing. Prof. Shah has been recognized by the National Science Foundation with a Faculty Early Career Development (CAREER) award and by MIT Technology Review on its 35 Innovators Under 35 list. She was also the recipient of the 2018 IEEE RAS Academic Early Career Award for contributions to human-robot collaboration and transition of results to real world application. She has received international recognition in the form of best paper awards and nominations from the ACM/IEEE International Conference on Human-Robot Interaction, the American Institute of Aeronautics and Astronautics, the Human Factors and Ergonomics Society, the International Conference on Automated Planning and Scheduling, and the International Symposium on Robotics. She earned degrees in aeronautics and astronautics and in autonomous systems from MIT and is co-author of the book, What to Expect When You're Expecting Robots: The Future of Human-Robot Collaboration (Basic Books, 2020).

Sarath Sreedharan (sarath.sreedharan@colostate.edu)

Sarath Sreedharan is an Assistant Professor at Colorado State University. His core research interests include designing human-aware decision-making systems that can generate behaviors that align with human expectations. He completed his PhD from Arizona State University, where his doctoral dissertation received one of the 2022 Dean's Dissertation Award for Ira A. Fulton Schools of Engineering. His research has been published in various premier research conferences, including

AAAI, ICAPS, IJCAI, AAMAS, IROS, HRI, ICRA, ICML and ICLR, and journals like AIJ. He has presented tutorials on his research at various forums and is the lead author of a Morgan Claypool monograph on explainable human-AI interactions. He was selected as a DARPA Riser Scholar for 2022, a Highlighted New Faculty by AAAI, and 10 to watch in AI by IEEE. His research has won multiple awards, including the Best System's Demo and Exhibit Award at ICAPS-20 and Best Paper Award at Bridging Planning & RL workshop at ICAPS 2022. He was also recognized as a AAAI-20 Outstanding Program Committee Member, Highlighted Reviewer at ICLR 22, IJCAI 2022 Distinguished Program Committee Member and Top Reviewer at NeurIPS 22.

SILVIA TULLI (silvia.tulli@sorbonne-universite.fr)

Silvia Tulli is an Associate Professor at Sorbonne University's Institute of Intelligent Systems and Robotics (ISIR). Her research focuses on explainable sequential decision making for AI agents, enabling them to bridge knowledge gaps by generating and reasoning about explanations of their internal processes, particularly when agents have different task models. She has co-organised workshops and seminars at AAAI (2021-2022), ICAPS (2021-2024), ICRA (2023), and IJCAI (2023), focusing on explainability in AI planning, agency, and robotics. She also co-organized a Dagstuhl Seminar on Explainable AI for Sequential Decision Making (2024).

PULKIT VERMA (pulkitv@mit.edu)

Pulkit Verma is an incoming Assistant Professor at Indian Institute of Technology Madras, India, and is currently a postdoc at MIT CSAIL, USA. His research focuses on the safe and reliable behavior of taskable AI agents. His research has been published at venues like NeurIPS, AAAI, AAMAS, KR, and ICAPS. He has organized multiple workshops including the two Workshops on Generalization in Planning at IJCAI 2022 and NeurIPS 2023, IJCAI 2025 Workshop and AAAI 2024 Spring Symposium on User-Aligned Assessment of Adaptive AI Systems, and ICAPS 2025 Workshop on Human-Aware and Explainable Planning. He has also been part of the program committee of multiple AAAI, NeurIPS, ICLR, IJCAI, ICAPS, R:SS, IROS, and AAMAS conferences.

6.2 PROGRAM COMMITTEE

We plan to keep the program committee large enough to ensure that each paper receives 3 reviews and no member reviews more than 3 papers. It is also diverse enough to ensure that the PC members will not be assigned papers that they have a conflict of interest with according to NeurIPS' definition of Conflict of Interest. Assignments of papers involving CoIs with an organizer will be handled by other members of the organizing team in addition to using OpenReview's automated CoI settings.

Members of the proposed program committee include:

- Nitay Alon, The Hebrew University, Israel.
- Kim Baraka, Vrije Universiteit (VU) Amsterdam, Netherlands
- Serena Booth, Brown University, USA
- Sonia Chernova, Georgia Tech, USA
- Ishita Dasgupta, Google DeepMind NYC, USA
- Harmen de Weerd, University of Groningen, Netherlands
- Sam Gershman, Harvard University, USA
- Elena L. Glassman, Harvard University, USA
- Matthew Gombolay, Georgia Tech, USA
- Noah Goodman, Stanford University, USA
- Alison Gopnik, UC Berkeley, USA
- Tom Griffiths, Princeton University, USA
- Mahdi Khoramshahi, Sorbonne University, France
- Brenden M Lake, New York University, USA
- Sheila McIlraith, University of Toronto, Canada.

- Francisco S. Melo, INESC-ID, Instituto Superior Técnico, Portugal
- Pierre-Yves Oudeyer, Inria Bordeaux, France
- Gözde Gül Şahin, Koç University, Türkiye
- Lindsay Sanneman, Arizona State University, USA
- Alberto Sardinha, PUC-Rio and GAIPS, INESC-ID, Brazil
- Miguel Vasco, KTH Royal Institute of Technology, Sweden
- Tan Zhi Xuan, National University of Singapore, Singapore