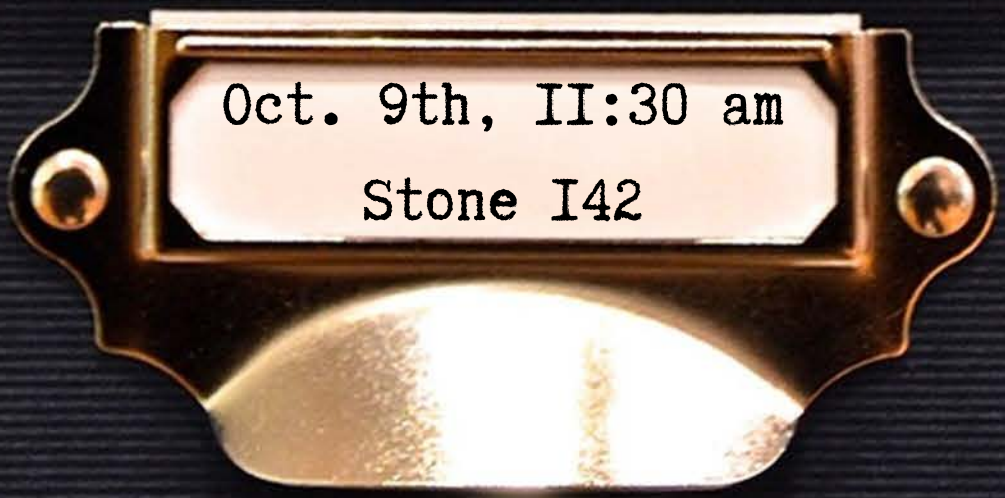
Opening the Black Box for Reinforcement Learning

UNCG Computer Science Colloquium Series

Recent advances in reinforcement learning have solved hard problems which were not able to solve previously such as Go, video games, self-driving cars, automation, financial and marketing decisions. Many of the successful models, however, are black box models that are hard to understand the knowledge discovered by the reinforcement learning agents. Therefore, when some catastrophic errors or misbehaviors are observed, there is no way to figure out why it occurs. As machine learning applications penetrate the market, interpretable models are significant to ensure the reliability of AI. What do we need to understand an AI? Can we collect some evidence to better understand it? What can we do if we have some evidence collected? These discussions will help us "open the black box."



Prof. Minwoo "Jake" Lee is an assistant professor in the Department of Computer Science at University of North Carolina, Charlotte. His main research interests are in the area of machine learning, with an emphasis on reinforcement learning. His recent publications discuss diverse topics in interpretable learning, knowledge representation and transfer, and applications in adaptive systems, natural language processing, brain-computer interfaces, robotics, and human-robot interactions. He received a Ph.D. from Colorado State University in May 2017.