G-LAMP SEMINAR

From Black-Box AI Workflows to Principled Human–AI Collaboration

Young Hyun Cho

Harvard University

일시: 2026/11/10 (화) 11:00-12:00
장소: 종합과학관 DB106 호

Abstract: Artificial intelligence (AI) is increasingly being used to perform tasks that were traditionally handled by humans. For organizations considering AI adoption, two questions are particularly important. How can they trust the quality of AI-generated outputs, and how much human work can AI safely automate? In this talk, I study these questions in the context of generate–verify workflows, where an AI generator produces answers and a verifier evaluates them, potentially over multiple rounds of revision. Providing statistical guarantees for such workflows is challenging because the generator and verifier are often black boxes, and later answers may depend on earlier feedback. A key observation is that organizations already have human experts who can evaluate AI-generated answers using the same standards they apply to their own work. We use these expert evaluations to develop a statistical framework for deciding when AI can safely handle tasks and when human intervention is needed. Our method provides finite-sample quality guarantees, and our theoretical results characterize how many expert evaluations are needed to achieve a desired level of automation. More broadly, this work illustrates how statistical thinking can help identify and address new problems arising from AI adoption. By recognizing that human expertise provides a source of reliable labels, we can turn a practical question about trusting AI into a statistical problem that can be studied using familiar principles of inference and uncertainty quantification.

주관 김세환 교수 (이화여자대학교 통계학과 / G-LAMP 사업단 LAMP 전임교원)

문의 02-3277-4089

주최 이화여자대학교 EWHA WOMANS UNIVERSITY · 인간중심인공지능연구원 Human-Centered Artificial Intelligence Research Institute · LAMP 대학기초연구소지원사업