AIから考える言語・知性・科学

大塚 淳（京都大学文学研究科）
竹内 孝（京都大学情報学研究科）
横井 祥（東北大学情報科学研究科）
包 含（京都大学白眉センター）

深層学習の科学的・工学的応用やテキストや画像などの生成AIの活用に見られるように、昨今の機械学習技術の進展は科学のあり方を大きく変えつつある。その一方で、機械学習モデルの理解可能性や信頼性については、まだ未知数なところが大きく、重要な解明課題となっている。また、複雑な認知タスクをこなす大規模言語モデルの登場は、そもそも言語とは、知性とは、そして科学とは何か、という哲学的な問題にも再考を促す。

このような問題を考えるための糸口として、本ワークショップでは「予測＋α」という切り口に焦点を当てる。現在の深層学習の成功は、その予測・汎化能力の高さに裏付けられている。昨今話題になっているChatGPTに代表される生成モデルも、基本的にはこうした予測のフレームワークの延長線上にあるといえよう。そうだとすると、機械学習の成功は、予測・汎化性能の向上に尽きるのだろうか？それとも、そこに回収されない「＋α」が必要とされるのだろうか、またそうだとしたらそれは何なのだろうか？とりわけ関心をひくのは、理解、概念操作、公平性、信頼性などといった、我々の言語的・知的・科学的活動を特徴付ける諸性質が、どの程度予測に帰着され、あるいはされないのか、という問題である。本ワークショップでは機械学習研究の最先端を行く研究者3名を迎え、「予測＋α」をめぐるこれらの問題群に対してそれぞれの専門的観点から話題を提供することを通じ、機械学習研究の現状とその哲学的含意を考察したい。

当日はまず、統計学の哲学を専門とする大塚が、ワークショップ全体の趣旨を説明した後、現在の機械学習、とりわけ表現学習（representation　learning）の知見が、我々の概念的理解にどのような含意を持つか、また逆に我々が理解できるような表現を深層学習モデルのうちに見出すことの利点と欠点について、問題提起を行う。

自然言語処理、とくに言語の表現学習を専門とする横井は、言語モデルの成功によって我々が改めて直面することになった「理解とは何か」という問いに関わるいくつかの論点を提供する。とくに、(1)　言語モデルが予測という負荷だけを通じて世界の理解にある程度到達している（ように見える）のはなぜなのか、(2) 言語モデルが世界を理解しているかどうかについて我々はどこまで理解可能なのか、というふたつの論点について、最近のいくつかの研究潮流を踏まえて検討する。

機械学習とその基礎づけである統計的学習理論を専門とする包は、近年急速に需要が高まっている trustworthy machine learning (TrustML)、すなわち予測性能が高いだけでなく公平性やロバスト性も高く信頼に足る機械学習技術について話題提起を行う。TrustML　の枠組みでは予測性能とロバスト性のトレードオフを取ることが求められるが、ときに予測性能の低下という代償を払ってまで得ようとしている「信頼性」とは何なのだろうか。機械学習が獲得を目指そうとしている法則の普遍性について、学習理論的な立場から再検討する。

機械学習とデータマイニングによる応用的データ解析を専門とする竹内は、機械学習によるデータ駆動型予測に内在するバイアスに関する話題、データ駆動型予測に基づく意思決定の信頼性に関する知見を紹介する。

以上の話題提供およびそれに続くディスカッションを通し、機械学習の発展がもたらす哲学的含意について、多角的な観点から議論を行う。