Part of the Symposium, “Is Deep Learning the Answer for Understanding Human Cognitive Dynamics?” at CogSci 2024

“Black-box tools” and “Brain-inspired modeling” provide complementary insights into the representations and neural dynamics underlying typical and atypical language processing

Kuperberg GR.

Box's famous assertion that "All models are wrong, but some are useful" has become a catchphrase among cognitive scientists. However, for our models to be useful, we must tailor them to our questions of interest. To do this, my lab is using two distinct modeling strategies to ask when, where, and how the brain infers meaning from language. First, we employ a diverse set of Natural Language Processing (NLP) models as theoretically motivated tools to probe distinct representations across the linguistic hierarchy during natural language processing. We integrate these models with time-sensitive neuroimaging techniques (EEG/MEG) to determine where and when the brain builds these representations during typical language comprehension. We are also leveraging this approach to characterize the disorganized speech patterns produced by some people with schizophrenia during natural language production (“positive thought disorder”).

Most "black box" NLP architectures, however, have little in common with the neurobiological and cognitive architecture of the human brain. Therefore, to understand (a) how information interacts across the linguistic/cortical hierarchy, and (b) how this gives rise to the dynamics of neural activity evoked by each word during real-time processing, we built a biologically and cognitively plausible, predictive coding model of lexico-semantic processing. This model is small-scale, with pre-specified representations at each level of its hierarchy, meaning that its dynamics are highly interpretable. Our simulations show that predictive coding (a) explains how information is propagated up and down the linguistic hierarchy, (b) predicts where these effects localize across the cortical hierarchy, (c) captures complex interactions between top-down contextual effects and bottom-up lexical effects on both neural activity and behavior, and (d) provides a natural, intuitive, and biologically plausible explanation for the time-course of both univariate and multivariate neural activity. Together, our "NLP models as tools" and "brain-inspired modeling" approaches are highly complementary, with each providing unique insights into both healthy and atypical language processing.