Centre - Variation - Multilingualism

Linguistic Lecture Series Summer Term 2026

Charles Rathkopf:

"How Should We Talk about the Mental States of AI Models?"

Abstract

We often describe LLMs in the same psychological terms we use to describe people. We say a model understands a question, refuses a request, follows an instruction, or tries to deceive. Can such claims ever be literally true? Or are they always false, merely metaphorical, or otherwise defective? I argue that they can be literally true, but that many current uses are nonetheless defective.

To accept that such claims can be literally true is to deny that linguistically articulate forms of cognition are uniquely human. That uniqueness claim has influential defenders. Some critics argue that psychological descriptions of LLMs are just anthropomorphic projections. Some go further and say that such claims are, for that very reason, irresponsible. I respond that mental properties need not come bundled in the package familiar from the human case. LLMs lack many important mental properties that humans have, but not all.

This leaves us with a harder question. If some psychological descriptions of LLMs can be literally true, why are they so often misleading? My answer is that ordinary psychological terms carry background assumptions inherited from the human case: about agency, stability, reciprocity, and accountability. Current systems do not support these assumptions. The problem is therefore not solved by choosing between literal truth and metaphor. We need new construals: shared ways of saying which parts of the human psychological picture apply to AI systems, and which do not. Claims about AI deception, where the construal problem is especially sharp, will serve as a test case.

SVM ZENTRUM Sprache | Variation Mehrsprachigkeit

Universität Potsdam