Theory of Mind and LLMs

- We are not aware of what actually happens in our own brains; moreover, experiments reveal that we often make decisions well before we think we do [2]
- Thinking about the future involves carrying out something like an inner dialogue, with an “inner storyteller” proposing ideas, in conversation with an “inner critic” taking the part of your future self. [2]
  - => We apply ToM on ourselves, and that is what we call consciousness
  - Some people also apply ToM on LLMs [5, 7], which might be due to the human nature [5]
- But if we assess our **unaccessible** internal states ourselves and say we are conscious and intelligent, how can we state for sure that the same thing is not going on in **unaccessible** internal states of LLMs? After all, if you ask an LLM what it is, it can define itself (apply ToM?)

=> Opinion: LLMs have already achieved some key aspects of meaning [4]

New Form of Intelligence?

- When applying tests designed for humans to LLMs, interpreting the results can rely on assumptions about human cognition that may not be true at all for these models => anthropocentric bias [1, 3]
  - => Opinion: intelligence and understanding “are the wrong categories” for talking about LLMs [1]
  - Open question: would it make sense to see the systems’ behavior not as “competence without comprehension” but as a new, non-human form of understanding? [1]
- Intelligent thought could be a mosaic of simple operations that, when studied up close, disappeared into its mechanical parts (c) Max Newman [2]
- „Real“ language understanding and intentionality consist of attributions of **unobservable** mental states [7] and it is unclear how we can meaningfully test for the „realness” of thoughts, feelings etc. [2]

=> Opinion: we need better measures for evaluating thought competence in LLMs before we can draw conclusions [7]