Internal States

- All said above makes us conclude there should be something more behind meaning than just reference
- Opinion: the interrelation of concepts is the key to meaning, and reference is just one optional aspect in world conceptualization [4]
  - When the associated terms shift in meaning, the resulting meaning adjusts [4]
  - Thus, a concept builds on conception of how the underlying pieces relate [4]
  - => the search for meaning should focus on understanding the way that the systems' internal representational states relate to each other [4]
- LLMs can manipulate complex concepts (but only in a linear manner) [3, 4]
- fMRI evidence supports that the human mental models for representational geometry are similar to LLMs [4] (?)
  - Cf. embedding and their geometrical meaning [7]
- LLM's internal state has some notions of conceptual role, so LLM's utterances have the semantic intent corresponding to these roles. [4]
- => Opinion: LLMs likely already share the foundation of how our own concepts get their meaning [4]

Theory of Mind and LLMs

- Theory of mind (ToM) is the ability to attribute mental states (emotions, intentions, knowledge) to oneself and others, and to understand how they affect behavior and communication.
  - When communicating with people, we build a partial **model** of who they are and what common ground they share with us, and use this to interpret their words [5]
  - Human language takes place between persons who share common ground [5]
- Opinion: just as LLMs can *model* a concept of color, shape etc. only from text, it can also share concepts with the human and succeed in ToM
  - Text provides such clues because humans generated the text [4]
  - When „talking“ to an LLM, it keeps record of its generated preferences, it tries to „understand“ your mood, intensions etc. [2]
  - Tests reveal modern LLMs have a high ToM [1, 3]