Universiteit
Leiden
The Netherlands

Theory of mind in language, minds, and machines: a multidisciplinary approach

Dijk, B.M.A. van

Citation
Dijk, B. M. A. van. (2025, January 17). *Theory of mind in language, minds, and machines: a multidisciplinary approach*. Retrieved from https://hdl.handle.net/1887/4176419

Version:	Publisher's Version
License:	Licence agreement concerning inclusion of doctoral thesis in the Institutional Repository of the University of Leiden
Downloaded from:	https://hdl.handle.net/1887/4176419

Note: To cite this publication please use the final published version (if applicable).

Summary

Typical human beings assume the perspective of other human beings frequently and with little effort. Doing so can be described as having a theory about what these other humans believe, desire, and intend, that is, as having a *Theory of Mind*. Having a Theory of Mind is an extremely useful tool for coordinating other humans' behaviours and navigating the social world. And just like this summary can be thought of as the linguistic code compressing the author's mind, so do humans encode and decode Theory of Mind often (but not exclusively) with spoken and written language.

Humans refine their Theory of Mind and language competences from a very young age. Since the two are linked, it is worthwhile to examine their intersections, particularly in *storytelling* as a phenomenon that engages both. Stories generally invite an audience to get immersed in a story world populated by story characters that have their own mental lives. Hence, stories are natural loci for trying to answer questions about Theory of Mind and language. For example, what do the character minds children create and their linguistic representations look like? What are more generally the 'linguistic fingerprints' of stories that deal with character minds of varying complexities? And, given that storytelling involves a Theory of Mind about the audience beyond the narrative, what are the properties of the storytelling language? This dissertation addresses these questions through the compilation of a Dutch children's story corpus on which various research methods are employed, varying from manual annotation to information extraction stemming from computer science.

Drawing on computer science here is no coincidence. Current artificially intelligent systems, in particular *large language models*, have become fluent language users to a degree of sophistication that was hard to foresee a decade ago. Hence, it is natural to ask in what way we can use these computational models to unravel the story language of children in novel ways. And to ask how adequately these large language models handle character minds, beliefs, desires and intentions in stories themselves

in comparison to children. And to ask what a philosophically sound way to better understand large language models' understanding can be. Though this dissertation started with employing techniques from computer science to extract information from stories, the latter questions also shift the role played by computer science - in particular language models - in this dissertation: from toolkit, to representation of mature language use, to subject in Theory of Mind experiments, and ultimately as focal point of a philosophical discussion revolving around better understanding machine understanding.

This dissertation has found various answers to these two different but complementary strands of questions. Regarding children, language and Theory of Mind, we found story characters with different kinds of mental lives. Story characters range from 'flat' in the sense of being merely staged or performing actions without clear goals, to more developed in featuring goal-directed behaviour and perceptions and emotions, to fully blown 'round' characters manifesting explicit and (complex) mental states. As characters become mentally more complex, stories display more complex linguistic profiles: they feature more complex syntax, more frequent use of pragmatic markers, a more complex and diverse vocabulary, and more explicitly linked clauses, among other things. Further, the language of live storytelling reflects the optimal balance between speaker and listener needs more closely than a comparable corpus of written stories, testifying to the social nature of live storytelling.

Regarding Theory of Mind and language in the context of modern artificial intelligence, we found that a language model can be used to disclose complex attentional and cognitive semantics in the way children use Dutch perception verb *zien* ('to see') in their stories. Yet, as subjects in Theory of Mind experiments, most language models do not fare better than children in dealing with beliefs, desires and intentions in stories, except for GPT-4, GPT-3.5, and PaLM2-chat. And in trying to better understand the kind of understanding that language models possess, we argued for a pragmatic framework where the aptitude of using 'understanding', 'believing', and 'thinking' and similar language for encoding the mental realm, is acceptable if this has practical value to us as humans and does not depend on the properties of the system.

By analysing Theory of Mind and language from a multidisciplinary angle, this dissertation aimed to contribute to an ongoing line of research that hopefully continues to entice both the theories and minds of researchers and the broader public.