



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).

Samenvatting

De meeste mensen kunnen regelmatig en bijna moeiteloos het perspectief van anderen aannemen. Dit kunnen we beschrijven als het hebben van een idee of theorie over wat een ander denkt, wil, of van plan is, oftewel over diens geest, vandaar dat deze vaardigheid ook wel *Theory of Mind* genoemd wordt. Het hebben van een Theory of Mind is een nuttig hulpmiddel om het gedrag van anderen te begrijpen en te voorspellen om zo de sociale wereld te kunnen navigeren. En zoals deze samenvatting gezien kan worden als de talige code die de gedachten van de auteur comprimeert, zo coderen en decoderen mensen vaak (maar niet uitsluitend) hun Theory of Mind met gesproken en geschreven taal.

Mensen verfijnen hun Theory of Mind en taalvaardigheden al vanaf jonge leeftijd. Aangezien de twee met elkaar verbonden zijn, is het de moeite waard hun raakpunten te onderzoeken in het *vertellen van verhalen*, een fenomeen dat uit beide vaardigheden put. Verhalen nodigen een publiek uit om zich onder te dompelen in een verhaalwereld die bevolkt wordt door personages met hun eigen mentale levens. Verhalen zijn dus uitermate geschikt voor het beantwoorden van vragen over Theory of Mind en taal. Welke soorten verhaalpersonages creëren kinderen, en hoe krijgen deze vorm in taal? Wat zijn in bredere zin de ‘taalkundige vingerafdrukken’ van verhalen die verschillende soorten verhaalpersonages opvoeren? En, aangezien verhalen vertellen een activiteit is die een Theory of Mind over wat er omgaat in het publiek vereist, wat zijn de eigenschappen van dit taalgebruik? Dit proefschrift beantwoordt deze vragen met een nieuw Nederlands kinderverhalencorpus dat ontrafeld wordt met methoden variërend van handmatige annotatie tot informatie-extractie uit de informatica.

Dat de informatica hier om de hoek komt kijken is geen toeval. Moderne systemen in de kunstmatige intelligentie, in het bijzonder de *large language models* oftewel grote taalmodellen, zijn vloeiende taalgebruikers geworden op een niveau dat een decennium geleden nog moeilijk te voorzien was. Het ligt daarom voor de hand

om te vragen op welke manier we deze computationele modellen kunnen gebruiken om de taal van kinderverhalen te analyseren. En om te vragen hoe goed deze grote taalmodellen zelf in staat zijn om te gaan met de overtuigingen, wensen en intenties van verhaalpersonages, in vergelijking met kinderen. En om te onderzoeken op welke manier we het begrijpen van taalmodellen moeten begrijpen. Waar dit proefschrift begint met het gebruiken van technieken uit de informatica om informatie uit kinderverhalen te extraheren, verschuiven deze vragen langzaam de rol die de informatica speelt in dit proefschrift - in het bijzonder de grote taalmodellen: van gereedschapskist, naar afspiegeling van volwassen taalgebruikers, naar proefpersoon in Theory of Mind-experimenten, tot brandpunt van een filosofische discussie over het beter begrijpen van het begrijpen van machines.

Dit proefschrift biedt de volgende antwoorden op deze twee verschillende maar complementaire stromen van vragen. Met betrekking tot kinderen, taal en Theory of Mind, bestaan er verhaalpersonages met verschillende soorten mentale levens. Verhaalpersonages variëren van 'platte' personages slechts aanwezig of handelend zonder duidelijke doelen, tot verder ontwikkelde personages met doelgerichte acties en (complexe) waarnemingen en emoties, tot volwaardige 'ronde' personages met expliciete (complexe) mentale toestanden. Naarmate personages mentaal complexer worden, vertonen verhalen ook vaker complexere taalkundige profielen. Zo bevatten ze vaker voegwoorden, complexere zinsstructuren, pragmatische markeringsen, en een complexer en diverser vocabulaire. Verder zijn in gesproken verhalen de behoeften van spreker en luisteraar beter in balans dan in geschreven verhalen, wat de sociale aard van 'live' verhalen vertellen in de klas benadrukt.

Met betrekking tot Theory of Mind, taal en moderne kunstmatige intelligentie, vonden we dat met een taalmodel kan worden geïllustreerd dat jonge kinderen het Nederlandse perceptiewerkwoord 'zien' al in de zin van 'begrijpen' of 'realiseren' kunnen gebruiken. Als proefpersonen in Theory of Mind-experimenten presteren de meeste taalmodellen echter slechter dan kinderen in het omgaan met informatie over overtuigingen en intenties in verhalen, met uitzondering van GPT-4, GPT-3.5 en PaLM2-chat. En in het beter begrijpen van het begrijpen van taalmodellen, wordt een pragmatisch kader geboden waarin het gebruik van mentale termen zoals 'begrijpen', 'geloven', en 'denken' acceptabel is zolang dit de mens praktisch dient; de fysieke eigenschappen van het onderliggende systeem spelen hier minder een rol.

Door deze multidisciplinaire benadering van Theory of Mind wil dit proefschrift bijdragen aan een onderzoekslijn die hopelijk zowel theorieën als geesten van onderzoekers en het bredere publiek blijft prikkelen.