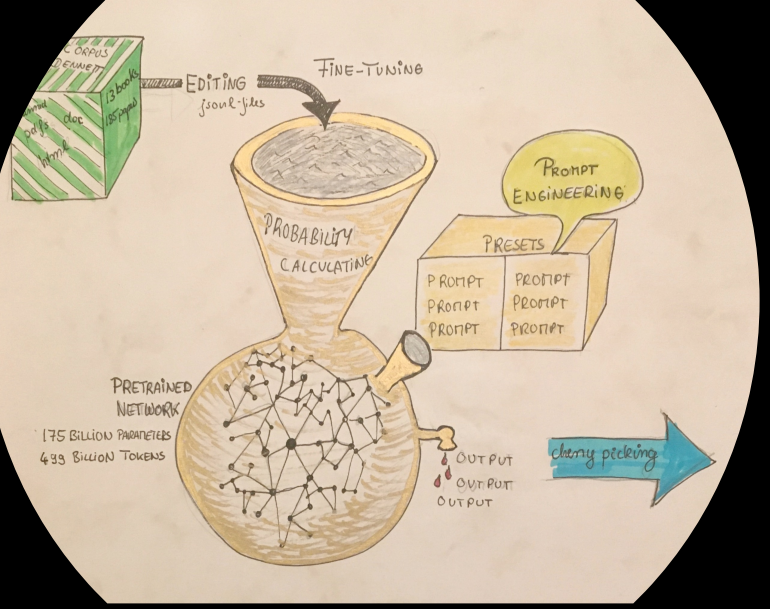
Hybrid Workshop on the Philosophy of Large Language Models
Eindhoven University of Technology

A new kind of comprehension in large language models?

ANNA STRASSER 2023

Preliminary note



Is philosophy safe from AI takeover?

DigiDan: a fine-tuned LLM on the corpus of Daniel Dennett
(Strasser, Schwitzgebel & Crosby 2022; Schwitzgebel 2022)

- difficult even for experts to distinguish text generated by Dennett from text generated by our model (Schwitzgebel et al. forthcoming)

(Results: The Computerized Philosopher)

I asked myself whether I was happy that I got involved in this project.

'Talking' to philosophers is much more attractive than only interpreting their textual outputs

An anecdote:
The night before my oral examination on Kant, I had this dream, in which I found myself discussing with Kant and even convincing him of something.

→ good dose of self-confidence for my exam
→ influenced my further work in philosophy by making me develop a strong preference to deal more with living than with already deceased philosophers

Blurring the difference between humans & machines

I hope no language model will ever be trained with all my statements.

- I do not aim to be mistaken for such a model.
- I do not aim to have such a digital legacy* continuing to make statements on my behalf after my death.

* For details, on what I mean by digital legacy see Karpus & Strasser (submitted). *Persons and their digital replicas.*

Denkwerkstatt Berlin

What is a GPT-3?

A NEURAL NETWORK TRAINED TO PREDICT THE NEXT LIKELY WORD IN A SEQUENCE

Pre-trained

- 499 billion tokens*
  (Common Crawl / WebText / Books / Wikipedia)

Generative

- can generate long sentences
- not just yes or no answers or simple sentences

Transformer

- calculating the probability of the next word appearing surrounded by the other ones

Generative Pretrained Transformer

- a 175 billion parameter language model which shows strong performance on many NLP tasks

*1 token = significant fractions of a word (on average 0,7 words per token)

Anna Strasser

DENKWERKSTATT BERLIN

Why think about the potential abilities of LLMs?

THINK BEFORE FURTHER TECHNOLOGICAL DEVELOPMENTS SURPRISE US WITH ALL KINDS OF DIGITAL TOOLS

- EVERLASTING SKEPTICISM REGARDING ANY ATTRIBUTIONS OF CONSCIOUSNESS, SENTIENCE, COMPREHENSION
  - concepts are not well-defined
  - no widely agreed-on solution to the so-called other-mind problem
- DESPITE THE LACK OF A SAFE EPISTEMOLOGICAL STANDPOINT
  - we many living beings are undeniably sentient, conscious, and comprehending
  - thinking about ascribing sentience, consciousness, and comprehension to non-living entities may shed some light on conditions we presuppose regarding living beings

→ investigate the extent to which we can appropriately characterize LLMs as having comprehension

Do LLMs present a paradigmatical case for non-living entities exhibiting a mode of comprehension?

– a mode that doesn't have all the features that human comprehension has –

Are LLMs mindless, or do they comprehend?

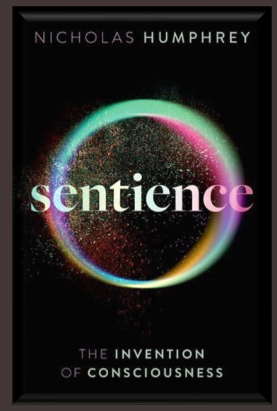
FIRST OBSERVATION

CONDITIONAL RELATIONSHIP BETWEEN CONSCIOUSNESS, SENTIENCE & COMPREHENSION IS UNCLEAR

- often treated as if they would always come in a package

THERE ARE REASONS TO QUESTION THIS

- Dennett: plants & bacteria are sentient but not conscious
  (Dennett as interviewed in Cukier 2022)
- Humphrey: one can have cognitive consciousness without phenomenal consciousness (sentience)



Not all sentient beings have what some philosophers call consciousness, and maybe even not all entities with comprehension are conscious or sentient.

contra machine comprehension

NO COMPREHENSION WITHOUT CONSCIOUSNESS

SHAKY EPISTEMOLOGICAL STANDPOINT

ANOTHER NECESSARY CRITERION

SKEPTICISM

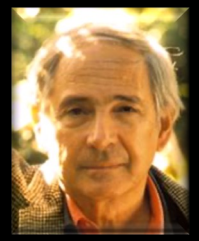
be more cautious about imputing comprehension

- humans tend to falsely impute comprehension due to the so-called Eliza effect
  - → imputing comprehension where there is no comprehension (*over-attribution*)

CONSCIOUSNESS = NECESSARY CONDITION

impossibility of conscious non-living entities:

- only living entities can have certain unspecified (mystical?) properties enabling experiences & mental states
- non-living entities lack those properties
  ..>question of comprehension does not arise (Searle 2010, p. 17)



➢ comprehension cannot be attributed to non-living entities

ONLY FULL-FLEDGED AUTONOMOUS AGENTS

- only entities that turn out to be agents with a high degree of autonomy have competence with comprehension
  - e.g., capable of revising their own selection processes to better achieve their goals (cherrypicking; Dennett forthcoming)

Anna Strasser

pro machine comprehension

But there are also voices claiming that there is "really" something we should call comprehension.
Some of them go as far as claiming that also sentience and consciousness are, in principle, conceivable.

IF sentience, consciousness, and comprehension do not necessarily come in a package

→ Under what circumstances are we inclined to ascribe comprehension to LLMs?
→ Is this of the same kind of comprehension we attribute to humans?
attribution of comprehension is not a matter of a dichotomous distinction but rather a matter of assumed gradualism

LOOKING AT LLMS

- mind-boggling outputs make it easy to suppose that there must be some comprehension
- often indistinguishable from human outputs (Strasser et al. 2022; Schwitzgebel 2022; Schwitzgebel et al. forthcoming)

➢ How should we characterize entities that are able to bring forward such outputs?

- humans able to produce comparable linguistic outputs are equipped with comprehension
- tempting to assume that comparable linguistic outputs from machines might also be accompanied by comprehension

But that may be a little premature conclusion because these might be cases of unjustified over-attributions.

A new kind of comprehension?

APPEARING FUNCTIONALLY EQUIVALENT TO HUMAN COMPREHENSION

- one of the many multiple realizations of one and the same ability

- realization of another ability or at least another mode of this ability

➢ must not be the same

→

How far can one get in arguing for the claim that LLMs might achieve a new kind of comprehension that is based on another mode of comprehension?

LLMs might achieve a new kind of comprehension

UNHUMAN-LIKE ERRORS

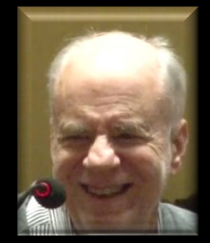
QUESTIONING WHETHER THE SUM OF OUTPUTS IS REALLY THE SAME

pointing to the *unhuman-like errors* of LLMs

- even though single outputs can be identical, the total of outputs is not the same illustration of unhuman-like errors
  - many examples of how easy it is to expose LLMs
  - outputs showing a lack of common sense

(cp. Mitchell & Krakauer 2022)

KRIPKE'S RULE-FOLLOWING



- we can never be sure whether a counterpart follows the same rules
  - the other could be "*quadding*" and not adding (Kripke 1982)

new kind of comprehension

- LLMs are '*qua-comprehending*' rather than 'comprehending'
  - ➢ not doing the same thing as humans when they produce linguistic output
- even if future LLMs make fewer *unhuman-like errors*, there still are possible outputs showing that the LLM followed another rule

LLMs might achieve another mode of comprehension

DIFFERENCES REGARDING THE VAST NUMBER OF MULTIPLE REALIZATIONS

MULTIPLE REALIZATIONS

human cognition

- limited regarding the amount & speed of data that can be processed
- what makes human cognition so impressive is that it requires so little data

➢ human comprehension is not based on the massive statistical models that today's LLMs use

- of course, humans are able to discover regularities & the totality of their experiences can have conditioning effects

abilities of LLMs

- based on a huge amount & speed of processed data

➢ constitute a different mode of the abilities we find in humans?

Differences in the way how outputs are realized can indicate a difference in the resulting abilities.

another kind of ability → might constitute a new kind of comprehension for which it might not be necessary to presuppose sentience & consciousness

* A possible objection could maintain that it is well possible that massive statistical models are also developed in humans, not over ontogenetic time scales, but phylogenetic/evolutionary time scales.

Possible objections

in favor of LLMs are comprehending in the same way humans do

HARD-CORE ADVOCATES OF MULTIPLE REALIZATIONS

- difference in how an entity produces outputs cannot call into question whether we are justified to ascribe the very same competence to the entity

❖ BUT this presupposes that the outputs are comparable in all respects

UNHUMAN-LIKE ERRORS WILL DISAPPEAR

- errors will be avoidable by scaling up

❖ BUT leaves it open to explain why this is to be expected

? syntax-semantics debate: semantics emerges (mystically) by itself if you put enough syntax in your models?

A MORE GENERAL OBJECTION

- pro similarity of machine & human comprehension
  - output is good enough regardless of obvious errors
  - humans make mistakes as well

❖ BUT those mistakes might be a different kind of mistake

A WEAKER CLAIM:

Ascribing a new kind of comprehension instead of claiming that machines would have the same kind of comprehension as we assume humans have

Do it your way

NOT ALL NECESSARY CONDITIONS FOR HUMAN COMPREHENSION ARE ALSO NECESSARY FOR MACHINES

some conditions for machines are not necessary for humans

some conditions for human comprehension are not necessary for machines

A WIDE SPECTRUM CAPTURED BY THE NOTION OF COMPREHENSION
different instances (modes) are connected via a family resemblance

LLMs may qualify for a new kind of comprehension

- by "shortcut learning"
  - relying on the ability to recognize & encode all possible correlations in data
  - requiring the ability to process a huge amount of data in a short time
  - some necessary conditions for human comprehension turn out not to be necessary for non-living entities

➢ machine comprehension might be possible without consciousness

PRESSING QUESTION

Is it sufficient to be a great pattern finder with extensive statistical abilities?

disjunctive conceptual scheme

ALLOWING DIFFERENT CONDITIONS FOR COMPREHENSION TO BE REQUIRED OF DIFFERENT ENTITIES

→ NEW PROBLEMS

DIFFICULT TO BE RECONCILED WITH A GENERAL IDEA OF GRADUALISM

@ humans: comprehension is a matter of degrees

- developmental psychology: comprehension develops gradually → presupposed conditions can vary in expression

DISJUNCTIVE CONCEPTUAL FRAMEWORK

- not all conditions come with a necessity
  - ➢ difficult to judge whether fulfilling one potential set of conditions is to be regarded as more or less comprehending

e.g., psychiatric diagnostic manuals (family resemblance & gradual variations)

- diagnosed with a mental disorder: having a certain number of symptoms of various severity
- Is it "worse" to have more symptoms in a weak expression than fewer symptoms in a strong expression? (Strasser 2021, p. 1947)

HOW TO CONNECT VARIOUS INSTANCES

- data-intensive machine comprehension without consciousness
- data-poor human comprehension with consciousness

BORDERLINE CASES

How can we distinguish competence with comprehension from competence without comprehension?

INTEGRATING NEW KINDS OF COMPREHENSION INTO OUR FRAMEWORK

- make sure that no competence without comprehension is captured by the expanded framework
- clarify how minimal comprehension is to be distinguished from no comprehension

Is comprehending language just a game as playing chess?

GAME-PLAYING AIs

- outperform human experts
- competency to play certain games
- knowing how
- "can make genuine moves within a board game" (Frankish 2022)

What can game-playing AIs do?

- able to follow the rules
- comprehend something about the game
- no deep understanding of the rules

➢ know-how it is sufficient to follow the rules obtusely

WHAT ABOUT A LANGUAGE GAME?

- following rules to produce linguistic output

genuine moves within language game → performance of speech acts

BUT regarding language, we tend to presuppose comprehension and consciousness for the ability to play along the rules

(Frankish 2022)

→ Are there speech acts that are not necessarily accompanied by consciousness but nevertheless by comprehension?

comprehension without consciousness

CAN WE ASCRIBE A NEW MODE OF COMPREHENSION TO LLMS?

gradualist point of view → avoiding dichotomic distinctions

- continuum between what LLMs do & what we do when we speak (Frankish 2022)

minimal speech acts

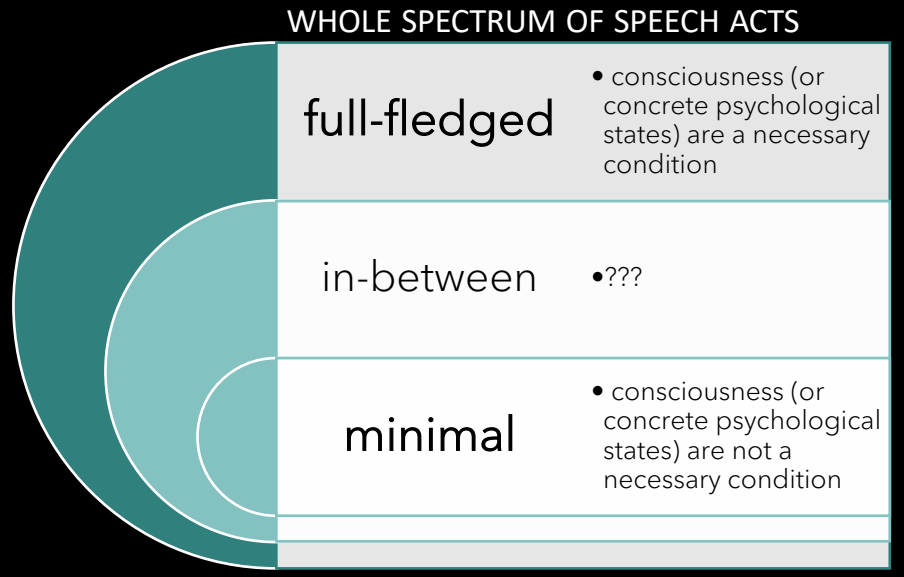
- training with large amounts of linguistic output
- gaining the ability to recognize regularities in language games
- deriving rules from the regularities

→ acting like an accomplished language player

→ appropriate contributions to conversations within limited domains

DERIVING 'RULES' → GAIN *KNOWLEDGE HOW*

REGARDING SOME PARTS OF THE LANGUAGE GAME HUMANS PLAY



COMPETENCE RESULTING FROM THIS *KNOWLEDGE HOW*

- just a competence without comprehension?

BUT

- a person has comprehension of a game when they have learned to follow the rules

→ LLMs gained comprehension

Let's call the competence to play along the rules the competence to produce minimal speech acts and allow this competence to be accompanied by a new kind of comprehension, what we could call minimal comprehension.

→ LLMs could become minimal language users by having the competence to produce certain kinds of speech acts

Humans can do a lot more with words

CAN WE REALLY CLAIM A CONTINUUM DESCRIBING DIFFERENT KINDS OF SPEECH ACT PERFORMERS?

→ reintroduce a sharp divide preventing us from de-psychologizing all types of speech acts

→ dichotomous distinction between minimal & full-fledged speech act performers

- minimal speech act performers just play along the rules to make appropriate contributions to conversations
- only full-fledged speech acts performers can use language as a tool

BUT

Can we exclude the possibility that we will discover grey areas in which non-living entities enter certain areas that we think are reserved for full-fledged speech act performers?

reintroduce a sharp divide
preventing us from de-psychologizing all types of speech acts

HUMANS & LANGUAGE-GAMES

(1) speech acts can overtake functional roles

- informing, instructing, persuading, encouraging, suggesting, deceiving …

→ presuppose psychological states

(2) speech acts presuppose having an understanding of their counterparts

- We don't play language games with stones since we are sure that they will not understand.

(3) often-emphasized abilities of humans

- they seem to be able to assess their own conversational contributions and select the best – they are prone to be cherry-pickers (Dennett forthcoming)

LLMs & LANGUAGE-GAMES

(1) speech acts are not based on psychological states

(2) it seems irrelevant whether others understand them

- LLMS play language games without presupposing that their outputs make sense for human interlocutors.
- Nevertheless, LLMs do not respond arbitrarily or completely randomly to linguistic input.
- Even if it is unlikely that machines have an idea of what human beings can comprehend, i.e., what a machine-generated linguistic output means for humans, they do have the ability to generate linguistic output that makes sense to humans.

(3) Interacting with LLMs, we often miss such cherry-picking features. Therefore, I would concede that at least recent LLMs only have a very limited ability to join our language games.

CONCLUSION

In some domains, LLM's performance is indistinguishable from human competencies

- LLMs = minimal speech act performers with a new kind of comprehension

In other domains, they seem merely to follow some rules obtusely and exhibit many *unhuman-like errors*

- competence without comprehension

- However, as long as we don't develop the ability to recognize whether they are just playing along the rules as minimal speech act performers, we should be careful …

Acknowledgements

This could have not happened without Eric and David Schwitzgebel!

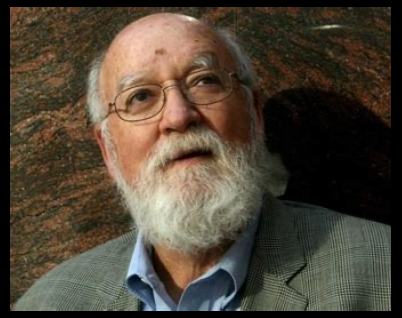
Special thanks to both Daniel C. Dennett & Matthew Crosby!

Dennett provided cooperation, advice, and encouragement in all aspects of this project.

Matthew Crosby provided technical expertise and implemented the fine-tunings for this project, as well as collaborating on a conceptual paper that provided the groundwork for this project (Strasser, Crosby, and Schwitzgebel forthcoming)

And last but not least, I want to thank Keith Frankish, Nick Humphrey, Joshua Rust and Manfred Frank for their thoughtful comments!

Humans and Smart Machines as Partners in Thought?



A hybrid workshop about large language models

UC RIVERSIDE

- hosted by the UC-Riverside Philosophy Department
- organized by Eric Schwitzgebel & Anna Strasser

SAVE THE DATE

10-11 May 2023

References

- Dennett, D. (forthcoming). We are all cherry-pickers.
- Cukier, K. (2022). Babbage: Could artificial intelligence become sentient? The Economist. https://shows.acast.com/theeconomistbabbage/episodes/babbage-could-artificial-intelligence-become-sentient
- Frankish, Keith (2022). Some thoughts on LLMs. *Blog post at* https://www.keithfrankish.com/2022/11/some-thoughts-on-llms
- Karpus, Jurgis & Strasser, Anna (submitted). Persons and their digital replicas.
- Krakauer, D. & Mitchell, M. (2022 - under submission as a Perspective article). The Debate Over Understanding in AI's Large Language Model. https://doi.org/10.48550/arXiv.2210.13966
- Kripke, Saul (1982). *Wittgenstein on Rules and Private Language: An Elementary Exposition.* Harvard University Press.
- Searle, John (2010). Why Dualism (and Materialism) Fail to Account for Consciousness. In Richard E. Lee (ed.), *Questioning Nineteenth Century Assumptions about Knowledge* (III: Dualism), New York: SUNY Press.
- Strasser, A. (2021). Fifty Shades of Social Cognition. How to Capture the Varieties of Socio-cognitive Abilities? *Proceedings of the Annual Meeting of the Cognitive Science Society*, 43.
- Strasser, A., Crosby, M. & Schwitzgebel, E. (2022). How far can we get in creating a digital replica of a philosopher? *Proceedings of Robophilosophy 2022. Series Frontiers of AI and Its Applications.* IOS Press, Amsterdam.
- Schwitzgebel, E. (2022). Results: The Computerized Philosopher: Can You Distinguish Daniel Dennett from a Computer? *Blog post at The Splintered Mind* (July 25) http://schwitzsplinters.blogspot.com/2022/07/results-computerized-philosopher-can.html
- Schwitzgebel et al. (forthcoming). Creating a Large Language Model of a Philosopher.