What we can learn from developmental psychology for dealing with non-understanding LLMs

Slides are downloadable at

Anna Strasser (2024)

Capacities of LLMs

WITH THE HYPE AROUND LLMS, EVERYONE SEEMS TO HAVE A STRONG OPINION ABOUT THE CAPACITIES OF LLMS

–

WHAT THEY CAN DO, CANNOT DO, MAY ONE DAY DO, AND WILL NEVER DO

ARTIFICIAL INTELLIGENCE | MAR. 1, 2023

You Are Not a Parrot

And a chatbot is not a human. And a linguist named Emily M. Bender is very worried what will happen when we forget this.

By Elizabeth Weil, a features writer at New York

OPINION

GPT-3, Bloviator: OpenAI's language generator has no idea what it's talking about

Tests show that the popular AI still has a poor grasp of reality.

By Gary Marcus & Ernest Davis

August 22, 2020

OpenAI

February 24, 2023

Planning for AGI and beyond

Our mission is to ensure that artificial general intelligence—AI systems that are generally smarter than humans—benefits all of humanity.

Blake Lemoine Follow

Jun 11 · 20 min read · Listen

Save

Is LaMDA Sentient? — an Interview

What follows is the "interview" I and a collaborator at Google conducted with LaMDA. Due to technical limitations the interview was conducted over several distinct chat sessions. We edited those sections together into a single whole and where edits were necessary for readability we edited our prompts but never LaMDA's responses. Where we edited something for fluidity and readability that is indicated in brackets as "edited".

(Bender et al., 2021; Open-AI ; Lemoine, 2022 ; Marcus & Davis, 2020; Weil, 2023)

Philosophical terminology is under discussion

Many terms that philosophers previously reserved for describing the distinguishing features of humans as rational agents are now being applied to machines, leading to intense debates over such notions as comprehension, knowledge, reasoning, and phenomenological consciousness.


Do Language Models Know When They're Hallucinating References?

Ayush Agrawal
Microsoft Research
t-agrawalay@microsoft.com

Mirac Suzgun
Stanford University
msuzgun@stanford.edu

Lester Mackey
Microsoft Research
lmackey@microsoft.com

Adam Tauman Kalai
OpenAI*
adam@kal.ai


Do Large Language Models Understand Us?

Blaise Agüera y Arcas


COGNITIVE SCIENCE
A Multidisciplinary Journal

Regular Article | Open Access

Do Large Language Models Know What Humans Know?

Sean Trott, Cameron Jones, Tyler Chang, James Michaelov, Benjamin Bergen

First published: 04 July 2023 | https://doi.org/10.1111/cogs.13309 | Citations: 1


Article

Human-like systematic generalization through a meta-learning neural network

WE SHOULD BE CAUTIOUS BEFORE CLAIMING STRONG OPINIONS!

Do LLMs understand us?

I think that LLMs do not understand us, but I also think that some interactions are strikingly similar to interactions among humans, which we call communication.

A widespread objection of describing interactions with LLMs as communication draws on their lack of understanding.

PROVOCATIVE QUESTION

IS UNDERSTANDING A NECESSARY CONDITION FOR ALL KINDS OF COMMUNICATION?

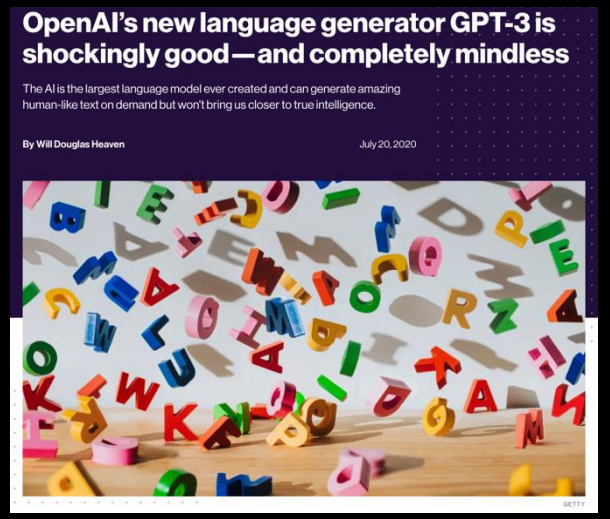
(Agrawal et al., 2023; y Arcas, 2022; Trott et al., 2023; Lake & Baroni, 2023)

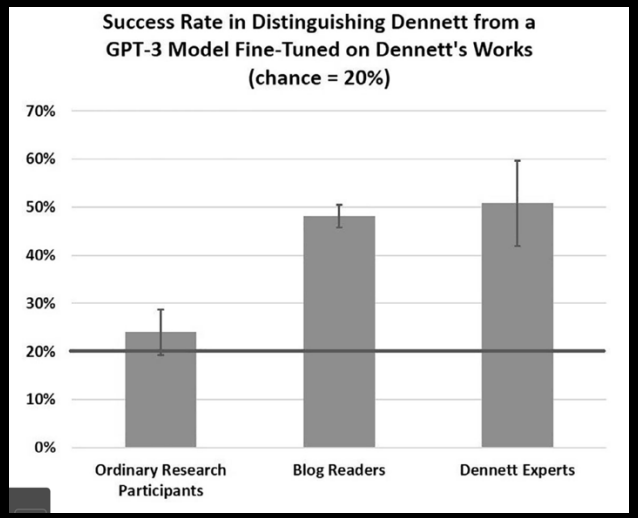
Amazing capacities

There is no question that LLMs have an amazing capacity to generate linguistic output that makes sense to humans!

notable successes in

- chess, go, discovering novel algorithms, protein folding *(Deep Blue, AlphaGo, AlphaTensor, AlphaFold)*
- automatic translation *(DeepL)*, lipreading *(LipNet)*
- computer code generation *(Github Copilot)*
- producing original prose with fluency equivalent to that of a human
  - ➢ **INDISTINGUISHABILITY**





(Campbell 2002; Silver et al. 2016, 2018; Ardila et al. 2019; Brown & Sandholm 2019; Jumper, Evans, & Pritzel et al. 2021; Fawzi et al. 2022; Assael et al. 2016; Steven & Iziev 2022; Heaven, 2020; Schwitzgebel et al., 2023)

DON'T LET YOURSELF BE CARRIED AWAY BY ALL THOSE AMAZING THINGS & FORGET TO NOTICE STRIKING DIFFERENCES

It is questionable whether they themselves can be said to understand what their outputs mean to us.

LLMs' outputs are not reliable; they hallucinate, they make severe mistakes …

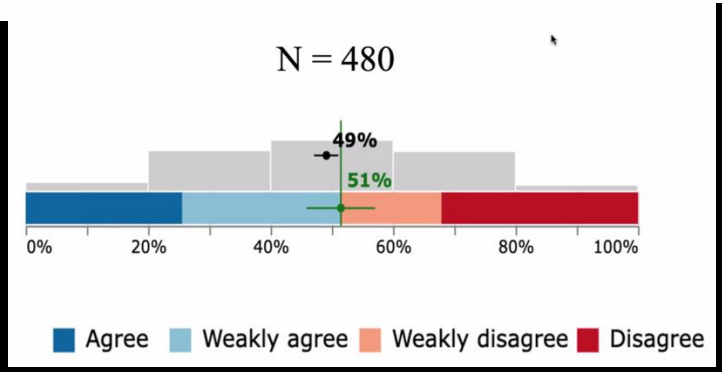
- they don't share a world with us
- ❖ they are not grounded
- ❖ they have no skin in the game
- ❖ they are not trained to consider the truth of utterances

WHAT DO NLP RESEARCHERS BELIEVE?
RESULTS OF THE NLP COMMUNITY METASURVEY

2022

Julian Michael[1,2] Ari Holtzman[1] Alicia Parrish[4] Aaron Mueller[5] Alex Wang[3]
Angelica Chen[2] Divyam Madaan[3] Nikita Nangia[2]
Richard Yuanzhe Pang[3] Jason Phang[2] and
Samuel R. Bowman[2,3,4]

Agree or disagree: Some generative models trained only on text, given enough data and computational resources, could understand natural language in some non-trivial sense.



Michael et al., 2022

Striking similarities

MOTHERBOARD
TECH BY VICE

'It's Hurting Like Hell': AI Companion Users Are In Crisis, Reporting Sudden Sexual Rejection

Replika, the "AI companion who cares," has undergone some abrupt changes to its erotic roleplay features, leaving many users confused and heartbroken.

By Samantha Cole

- 2023

Replika users feel like losing their best friend after an update

Nevertheless, humans do interact with these machines in ways that strongly resemble genuine conversation, and we need to find illuminating ways of describing this activity.

I THINK WE ARE RIGHT TO BE CONFUSED ABOUT THE CAPACITIES OF LLMS.

It is important that we elaborate on both the similarities and the differences, or, as I will frame it later, we should pay attention to the asymmetric features of HMIs.

My starting point

WHAT ARE WE DOING WHEN WE INTERACT WITH LLMS?

WE CAN NOT REDUCE ALL OF OUR INTERACTIONS WITH LLMS TO MERE TOOL USE

- Is an LLM or a robot developed with generative AI technology a person or a thing? → neither nor

BUT, so far, we have no philosophical terminology to describe what it is instead!

→ rethink our conceptual framework, which so clearly distinguishes between tools as inanimate things and humans as social, rational, and moral interaction partners

We need a conceptual framework that can capture
INBETWEEN PHENOMENA

(Strasser & Schwitzgebel 2024; Strasser, 2025)

Extreme positions

Hard-core instrumental view

NON-LIVING THINGS CAN NEITHER HAVE MORAL AGENCY NOR MORAL PATIENCY

In expectation of AGI view

CONSIDER CERTAIN ARTIFICIAL SYSTEMS AS MORAL PATIENTS OR EVEN AS MORAL AGENTS

PHILOSOPHY POSES TOO DEMANDING CONDITIONS → abilities of children, non-human animals, and artificial systems fall through the conceptual net

sophisticated terminology of philosophy prevents us from grasping the INBETWEEN

→ conceptual frameworks that can distinguish more finely-grained instances across a wider spectrum
→ capture phenomena one finds in developmental psychology, animal cognition, and AI

ARE LLMS OUT OF SOCIAL GAMES WHEN WE ARE CONVINCED THAT THEY CANNOT UNDERSTAND (COMPREHEND) LINGUISTIC OUTPUTS AS WE HUMANS DO?

To make progress here, I suggest

- **taking a closer look at the difference between *competence with comprehension* and *competence without comprehension***
- **asking if there are forms of communication for which a level of competence without comprehension is sufficient**

To this end, I shall look at the linguistic development of children and at other communicative situations where it is not obvious that both partners possess comprehension.

Even in interaction between humans, certain communicative activities, or language games, are asymmetric in the distribution of abilities.

To what extent do such language games offer a helpful template for describing human interactions with LLMs?

Not all things come in a package!

Might LLMs represent a paradigmatic case of non-living entities exhibiting a mode of understanding?
A mode that does not exhibit all the features of human understanding, especially not the feature of being conscious or sentient?

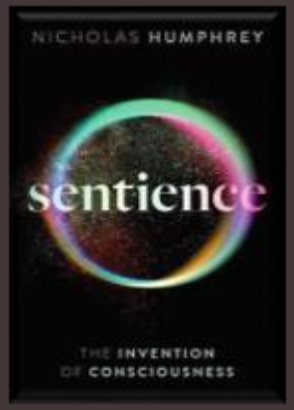
I do not think that we have reasons to ascribe understanding to LLMs!
Now, I wonder if LLMs need to understand at all in order to act as communication partners.

CONDITIONAL RELATIONSHIP BETWEEN DIVERSE MENTAL ASCRIPTIONS IS UNCLEAR

often treated as if they would always come in a package

NOT ALL THINGS COME IN A PACKAGE!

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



Does being a partner in a communicational setting always presupposes understanding in both involved partners?

Dennett & the four creatures

	implementation	properties	comprehension	learning
Darwinian	hard-wired	*clueless towards novel variations*	born knowing (gifted) **no comprehension**	learn nothing
Skinnerian	hard-wired • favor whatever has reinforcing outcomes	*some plasticity in a repertoire of behavior*	without knowing why they favor this **no comprehension**	learn • by trial-and-error
Popperian	free-floating maxim • look before leap • favor pretesting	*information sensitive & forward-looking processes*	without understanding why they engages in this pretesting **no comprehension**	learn • by testing candidates for action against information about the world stored in their brains
Gregorian	deliberately use thinking tools	*apply lessons to new material, new topics*	understanding the grounds of their own understanding **with comprehension**	lots of learning • improves generators & testers

LLMs DO NOT UNDERSTAND

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 (*We are all cherry pickers;* Dennett 2024)

Since I am convinced that artificial systems still do not yet qualify as full-fledged autonomous agents, even though I would ascribe minimal joint-action abilities in quasi-social interactions with humans to them, I am motivated to investigate whether all communicative settings presuppose comprehension of both participants-

So far, I have not given any working definition of understanding.

Some provisional things

1. I am using comprehension & understanding interchangeably
2. Dennett's distinction between the four creatures leads to a very demanding notion of comprehension, something very deep ...

BUT

- he also talks of sorta of comprehension
- He emphasizes that he is fond of a gradual approach

UNDERSTANDING IS NOT AN ALL-OR-NOTHING QUESTION

IMAGINE ALL KINDS OF LINGUISTIC INTERCHANGES YOU HAVE HAD IN YOUR LIFE

No communication

- interactant just talk past each other
  - nothing more than two entities taking turns in speaking

Successful communication

- real exchange in which both interaction partners understand each other
  - all kinds of speech acts are part of those interactions, people inform, warn, explain stuff to each other

CASES THAT FALL INTO NEITHER CATEGORY

- in which we can speak of more or less successful communication
- which have an **asymmetrical aspect**

THE INBETWEEN

NO COMMUNICATION
no comprehension
TOOL USE

talking past each other

SUCCESSFUL COMMUNICATION
lots of comprehension
FULL-FLEDGED SOCIAL INTERACTION

reciprocal understanding

ASYMMETRIC INBETWEEN PHENOMENA

only one agent has competence
with comprehension

Talking with kids

THE MOVE FROM COMPETENCE WITH SORTA COMPREHENSION TO FULL-FLEDGED COMPREHENSION IS A GRADUAL ONE

COMPLEX SOCIAL SKILLS DO NOT EMERGE IN AN INSTANT!

- **not developmentally in humans**
- **not phylogenetically in animal evolution**
- **not technologically in the design of AI systems**

- children frequently use words without understanding them
  - just repeating something they have heard before

BUT

- through repeated series of interactions in which they can observe the reactions of their interactants, they start to understand the meaning more and more

clearly categorize communication with children as communication, even if not always super successful

a gradualist approach → communication in its initial phases can be described as an asymmetrical interaction

- asymmetric:
  sets of conditions that have to be fulfilled by the interaction partners differ between children and adults

Towards asymmetric quasi-social interactions

NO NECESSITY OF AN EQUAL DISTRIBUTION OF ABILITIES AMONG ALL PARTICIPANTS

A HUMAN FEATURE

DEVELOPMENTAL PSYCHOLOGY

- joint action of adults and children
- children = socially interacting beings

ARTIFICIAL INTELLIGENCE

- quasi-social interactions of human beings & artificial systems
- artificial systems =?= quasi-socially interacting entities

ADULT & CHILD

ROBOT & HUMAN
LLM & HUMAN

ASYMMETRIC DISTRIBUTION OF CONDITIONS

By talking to children as if they would understand everything, we give them a chance to gain more and more understanding.

BUT

current AI systems are not capable of learning from our interactions with them in the way children do

- might become true for future machines, then we might say that treating them as social partners may help them develop the pattern of reactions that make them social partners.

→ we can refer to cases of communication in our everyday experience in which the communication partner lacks a great amount of understanding

More examples

Talking with very drunk people

Examinations with nervous students

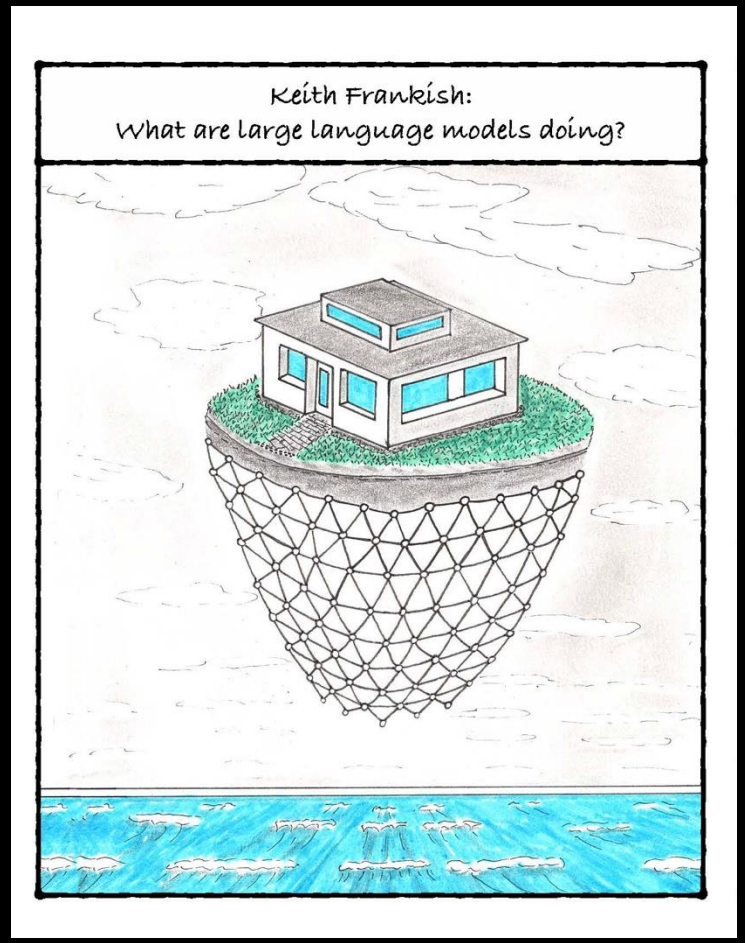
Bullshitting

Small talk

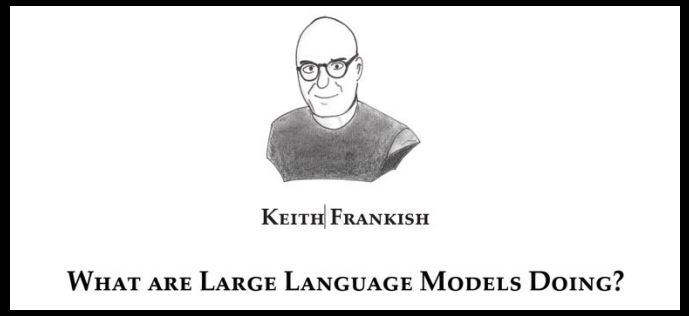
Not all language games we entertain presuppose full-fledged comprehension.

Somehow, it seems sufficient to follow some ‘easy’ rules, repeat patterns we have observed beforehand, and play all kinds of chat games.

Chat games



Artist: Moritz Strasser



- examines whether we should think of LLMs as capable of performing intentional actions guided by reasons and, more specifically, by communicative intentions.

➢ He argues that current LLMs are simply making moves in a narrowly defined language game (the “chat game”), and he suggests that LLMs’ responses are motivated solely by a desire to play this game and not by any communicative intentions.

MAKING MOVES IN A NARROWLY DEFINED LANGUAGE GAME CAN BE DONE WITHOUT MUCH COMPREHENSION

THERE ARE

COMMUNICATIVE ACTIVITIES, OR LANGUAGE GAMES,

THAT ARE ASYMMETRIC IN THE DISTRIBUTION OF ABILITIES.

Instead of a conclusion, I would like to pose the question of to what extent such language games offer a helpful template for describing human interactions with LLMs.

Are LLMs

- like children who never grow up
- like drunk communication partners who never become sober
- like skilled language game players who only have competence without comprehension?

All this work would not have been possible if I had not interacted with a lot of people & machines

Daniel Dennett

Eric Schwitzgebel

Mathew Crosby

David Schwitzgebel

Mike Wilby

DigiDan

Thank you !

References

Agrawal, A., Mackey, L., & Kalai, A. T. (2023). *Do Language Models Know When They're Hallucinating References?* (arXiv:2305.18248). arXiv. http://arxiv.org/abs/2305.18248

Agüera y Arcas, B. (2022). Do Large Language Models Understand Us? *Daedalus*, *151*(2), 183–197. https://doi.org/10.1162/daed_a_01909

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 610–623. https://doi.org/10.1145/3442188.3445922

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., … Amodei, D. (2020). *Language Models are Few-Shot Learners*. https://doi.org/10.48550/arXiv.2005.14165

Dennett, D. C. (2018). *From Bacteria to Bach and Back: The Evolution of Minds* (Reprint edition). W. W. Norton & Company.

Fawzi, A., Balog, M., Huang, A., Hubert, T., Romera-Paredes, B., Barekatain, M., Novikov, A., R. Ruiz, F. J., Schrittwieser, J., Swirszcz, G., Silver, D., Hassabis, D., & Kohli, P. (2022). Discovering faster matrix multiplication algorithms with reinforcement learning. *Nature*, *610*, 47–53. https://doi.org/10.1038/s41586-022-05172-4

Heaven, W. D. (2020). *OpenAI's new language generator GPT-3 is shockingly good—And completely mindless*. MIT Technology Review. https://www.technologyreview.com/2020/07/20/1005454/openai-machine-learning-language-generator-gpt-3-nlp/

Heaven, W. D. (2022, November 30). *ChatGPT is OpenAI's latest fix for GPT-3. It's slick but still spews nonsense*. MIT Technology Review. https://www.technologyreview.com/2022/11/30/1063878/openai-still-fixing-gpt3-ai-large-language-model/

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S. A. A., Ballard, A. J., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J., … Hassabis, D. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature*, *596*, 583–589. https://doi.org/10.1038/s41586-021-03819-2

Lake, B. M., & Baroni, M. (2023). Human-like Systematic Generalization through a Meta-learning Neural Network. *Nature*, 1–7. https://doi.org/10.1038/s41586-023-06668-3

Lemoine, B. (2022, June 11). Is LaMDA Sentient? — An Interview. *Medium*. https://cajundiscordian.medium.com/is-lamda-sentient-an-interview-ea64d916d917

Lemoine, B. (2023, February 27). I worked on Google's AI. My fears are coming true. *Newsweek*. https://www.newsweek.com/google-ai-blake-lemoine-bing-chatbot-sentient-1783340

Marcus, G., & Davis, E. (2020). GPT-3, Bloviator: OpenAI's language generator has no idea what it's talking about. *MIT Technology Review*. https://www.technologyreview. com/2020/08/22/1007539/gpt3-openai-language-generator-artificial-intelligence-ai-opinion

Michael, J., Holtzman, A., Parrish, A., Mueller, A., Wang, A., Chen, A., Madaan, D., Nangia, N., Pang, R. Y., Phang, J., & Bowman, S. R. (2022). *What Do NLP Researchers Believe? Results of the NLP Community Metasurvey* (arXiv:2208.12852). arXiv. https://doi.org/10.48550/arXiv.2208.12852

References

Open-AI. (n.d.). *Planning for AGI and beyond*. Retrieved 6 May 2024, from https://openai.com/index/planning-for-agi-and-beyond

Schwitzgebel, E., Schwitzgebel, D., & Strasser, A. (2023). Creating a large language model of a philosopher. *Mind & Language*, 1–23. https://doi.org/10.1111/mila.12466

Strasser, A. (2025). *Inbetweenism. Why ethical positions appear outdated in the face of the new AI technology*. De Gruyter.

Trott, S., Jones, C., Chang, T., Michaelov, J., & Bergen, B. (2023). Do Large Language Models Know What Humans Know? *Cognitive Science*, *47*(7), e13309. https://doi.org/10.1111/cogs.13309

Weil, E. (2023, March 1). *You Are Not a Parrot*. Intelligencer. https://nymag.com/intelligencer/article/ai-artificial-intelligence-chatbots-emily-m-bender.html

Finding our way through the jungle

TOOL KIT 'MINIMAL APPROACHES'

How to conceptualize phenomena in the field of developmental psychology & animal cognition that fall through the sophisticated conceptual net of philosophy

- questioning the necessity of far too demanding conditions
- considering multiple realizations of capacities that seemed to be restricted to sophisticated adult humans

Stephen Butterfill & Ian Apperly (2013): minimal mindreading | John Michael et al. (2016): minimal sense of Commitment | Elisabeth Pacherie (2013): shared intention lite
Anna Strasser (2006): minimal action