Can Large Language Models Think?

David Chalmers

APA Eastern Division Presidential Address 2023





Adventures in technophilosophy

VIRTUAL WORLDS AND THE PROBLEMS OF PHILOSOPHY



DAVID J. CHALMERS

The New York Times

Google Sidelines Engineer Who Claims Its A.I. Is Sentient

Blake Lemoine, the engineer, says that Google's language model has a soul. The company disagrees.



Student uses ChatGPT to write an essay for philosophy class in South Carolina – professor warns of 'flood' of chatbot scams as Al improves

Asson Regan December 27, 2022

Language Models

- A language model (LM) is a system that assigns probabilities to sequences of text, thereby predicting and generating text completions.
 - Input: "The quick brown fox jumps over the lazy"
 - Output: "dog" (99+%).

Large Language Models

- Large language models (LLMs) are giant artificial neural networks that serve as language models
 - typically arranged in a transformer architecture





Large Language Models

- Large language models (LLMs) are giant artificial neural network language models
 - typically using a transformer architecture
 - trained on data from all over the internet
 - up to 500 billion+ parameters
 - BERT, GPT-2, GPT-3, PaLM, ChatGPT, [GPT-4?], ...

Ethics/Value Questions

- Are LLMs safe?
- Are LLMs fair?
- Are LLMs reliable?
- Are LLMs explainable?
- Could LLMs have moral status?



Philosophy of Mind Questions

- Can LLMs be conscious?
- Can LLMs think?
- Can LLMs be agents?

Can LLMs Think?

- Can LLMs have genuine propositional attitudes such as beliefs, desires, judgments that p?
- Can they possess concepts?
- Can they understand language?

\$

There is no definitive answer to this question, as the concept of "thought" is difficult to define and there is no consensus on what it entails. Some people might argue that a large language model like Assistant can think, while others might disagree.

Here are some reasons that could be given for why a large language model might be able to think:

- 1. Language models are able to process and generate human-like text, which is a key component of human thought.
- 2. Language models can learn from data and make predictions based on that learning, which is similar to how humans use past experiences to inform their decision making.
- Language models can be trained on a wide range of tasks, including language translation, question answering, and image recognition, which suggests that they have a certain degree of flexibility and adaptability.
- 4. Some researchers have used language models to generate creative outputs, such as poetry and music, which could be seen as a form of "thought."

On the other hand, here are some reasons that could be given for why a large language model might not be able to think:

- Language models do not have a physical body or senses, and do not experience the world in the same way that humans do. This means that they do not have access to the same types of sensory input that humans do, and cannot engage in physical actions or movements.
- 2. Language models do not have consciousness or self-awareness, and do not have subjective experiences or feelings.
- Language models are not able to think independently, and can only produce outputs based on the data they have been trained on. They do not have the ability to form original ideas or perspectives.

Overall, whether or not a large language model like Assistant can be said to "think" is a matter of debate and may depend on how one defines thought.

The Turing Test

- Alan Turing (1950): A machine thinks iff it's indistinguishable in conversation from a human.
- VOL. LIX. No. 236.]

[October, 1950

MIND a quarterly review of psychology and philosophy

- LLMs don't quite pass the Turing test, but they're getting close.
- TT isn't conclusive proof of thinking, but it makes a defeasible case that needs defeating.

I:--COMPUTING MACHINERY AND INTELLIGENCE

By A.M. Tubing

Arguments Against Thinking in LLMs

- I'll focus on a certain class of arguments that it's impossible for an LLM to think.
- Roughly: arguments that LLMs lack the sensory grounding required for genuine thought, meaning, and understanding.

What I'll Do

- I'll analyze three such arguments, and argue for a "glass half-full" view:
 - These arguments don't establish that LLMs can't think, mean, or understand
 - They do establish some serious limitations, which might be overcome in extensions of LLMs
 - There's a prima facie case that in principle LLMs can think.

What I Won't Do

- Argue that AI is possible.
- Argue that current LLMs can think.
- Argue that LLMs can be conscious or can be agents.
- Address objections to LLM thinking from consciousness or agency.

Outline

- I. Arguments that thinking requires sensing.
- 2. Arguments from the poverty of text.
- 3. Arguments from disembodiment and knowhow.
- 4. Limitations of current LLMs.

Part I: Does Thinking Require Sensing?

- For an AI system to have genuine thought, meaning, and understanding, its processes must be appropriately grounded in senses and in the environment.
- LLMs lack appropriate grounding in senses and the environment, and so lack genuine thought, meaning, and understanding.

The Symbol Grounding Problem

 Stevan Harnad, "The Symbol Grounding Problem" (1990):

"To be grounded [needed to have genuine thought, meaning and understanding], the symbol system would have to be augmented with nonsymbolic, sensorimotor capacities -- the capacity to interact autonomously with that world of objects, events, actions, properties and states." Paul Vogt Yuuga Sugita Elio Tuci Chrystopher Nehaniv (Eds.)

Symbol Grounding and Beyond

Third International Workshop on the Emergence and Evolution of Linguistic Communication, EELC 2006 Rome, Italy, September/October 2006, Proceedings



🖄 Springer

Grounding Proponents

- Emily Bender and Alexander Koller (2020), "Meaning, form, and understanding in the age of data"
- Emily Bender, Timnit Gebru, et al (2021): "On the dangers of stochastic parrots"
- Brenden Lake and Gregory Murphy (2021), "Word meanings in minds and machines"
- Yann LeCun and Jake Browning (2022), "AI and the limits of language"

The "Chinese Room"

- Sometimes the symbol grounding problem is presented as a version of the "Chinese room" argument.
- I'll set aside general anti-Al arguments like this one.



The "Klingon Room"

939 INVIAL'STR SET 199638 959 YENTE'L STR TT TTY NOLL STR STR STR TTY TIS IND IN ALY STR STRAL INTERIAL INVIAL

Argument from Sensory Capacities

- I. LLMs lack sensory capacities.
- 2. Genuine thought requires sensory capacities.
- 3. So: LLMs lack genuine thought.

Reply to Premise I [LLMs Lack Senses]

- Standard LLMs have text input and output which is arguably a sort of sensory grounding.
- There are extended LLMs e.g. visionlanguage-action models — that have sensory grounding.





Figure 1: LLMs have not interacted with their environment and observed the outcome of their responses, and thus are not grounded in the world. SayCan grounds LLMs via value functions of pretrained skills, allowing them to execute real-world, abstract, long-horizon commands on robots.

Reply to Premise 2 [Thinking Requires Senses]

- To think about LLM capacities as well as Al capacities more generally, it's useful to think about the capacities of pure Al systems without grounding.
- Does genuine thought require sensory processes?

Sensory Grounding in History of Philosophy

- "The soul never thinks without an image."
 Aristotle.
- "There's nothing in the intellect that wasn't previously in the senses."—Aquinas.
- "All our simple ideas in their first appearance are derived from simple impressions." —Hume.
- These empiricist theses all tend to suggest that thinking requires [having had] senses.

Rationalism and Sensory Grounding

- Rationalists tend to deny that thought requires sensory grounding:
 - Plato: we have thought (about the forms) before we have senses and a body.
 - Descartes: the pure intellect thinks independently of the senses.

 For help with the history, thanks to: Peter Adamson, Max Cappuccio, Victor Caston, Becko Copenhaver, Christian Coseru, Keota Fields, Don Garrett, Sophie Grace, Steven Horst, Anne Jacobson, Anja Jauernig, Chad Kidd, Jonathan Kramnick, James Kreines, Béatrice Longuenesse, Jake McNulty, Jessica Moss, Elliot Paul, Lewis Powell, Naomi Scheman, Tobias Schlicht, Eric Schliesser, Lisa Shapiro, Karsten Struhl, Christina Van Dyke, Charles Wolfe, ...

The Sense-Thought Thesis

- Sense-thought thesis: Thought requires (having had) senses.
 [Necessarily, if x thinks at t, x has had the capacity to sense at or before t.]
- Scope: restricted to humans, or unrestricted.
 - Restricted thesis: focus of historical debate?
 - Unrestricted thesis: relevant to Al!

Sensing and Thinking

• Senses: capacity for sensory inputs, sensory processes, or sensory experiences.

[Visual imagery counts as sensory; introspection of thoughts doesn't.]

• Thinking: minimally, the ability to judge that p.

Part Ia: Pure Thinkers

 Key thought experiment: a "pure thinker" who has thought without sensory capacities.

Avicenna's Floating Man



Philosophy thought experiments in drawings, by @HelenDeCruz
Avicenna's Floating Man [Ibn Sina, De Anima, ~1027]

"...He was just created at a stroke, fully developed and perfectly formed but with his vision shrouded from perceiving all external objects – created floating in the air or in the space, not buffeted by any perceptible current of the air that supports him, his limbs separated and kept out of contact with one another, so that they do not feel each other. Then let the subject consider whether he would affirm the existence of his self. There is no doubt that he would affirm his own existence, although not affirming the reality of any of his limbs or any external thing."



Avicenna Abu-`Aly al-Husayn Ibn `Abdallah Ibn Sina



Metaphysics and Epistemology

- Avicenna draws a metaphysical conclusion: the floating man is aware of the self but not aware of the body, so the self isn't the body.
- Others draw epistemological conclusions about cognitive capacities:
 - E.g. Matthew of Aquasparta: floating man suggests that self-consciousness doesn't require sensory knowledge. The Fate of the Flying Man: Medieval

Reception of Avicenna's Thought Experiment

🐊 Juhana Toivanen

Floating Man: Imperfections

- Avicenna's floating man
 - has (unused) senses;
 - is a human (and a man).
- We need a generic label for a (possibly nonhuman) thinker without sensory capacities.
- I'll use "pure thinker".

Questions

- I. Is a pure thinker possible?
- 2. What could a pure thinker think?
- 3. What could a pure thinker know?

Is a Pure Thinker Conceivable?

- Prima facie, a pure thinker (thinking about mathematics, say, or executing the cogito) is conceivable.
- N.B. Not necessarily a human: arguably there has never been a human pure thinker.
- But there's no obvious contradiction in conceiving an AI system without senses (e.g. Sawyer, Wake), or a disembodied mind.

From Conceivability to Possibility?

- What might defeat the prima facie conceivability as a guide to possibility?
 - thought = human-style thought
 - human-style thought involves the senses
 - conceivable AI thought = schmought
- Doubtful that "thought" works this way; also, AI schmought seems good enough!

Strong Empiricism, Externalism, ...

- Possible defeaters: *all* thought is constituted by
 - sensory experience (strong empiricism)
 - environmental relations (strong externalism)
 - extended processes (strong extended mind)
 - embodied processes (strong embodied mind)
- Trouble: weak version doesn't suffice, strong version isn't plausible



■ The Oxford Handbook *of* 4E COGNITION

Strong Empiricism

- Strong empiricists may exist (Hume, Barsalou?), but the view is widely rejected.
- Even many empiricists reject the view for non-humans (e.g. Aristotle and Berkeley for God, Aquinas and Reid for souls or angels)
- The leading accounts of logical and mathematical concepts (even in humans) are non-empiricist.

Strong Externalism

- Standard externalist (extended, embodied views) hold that some concepts/thoughts are externally grounded.
- Strong externalism (...): all concepts/ thoughts are externally grounded
- ST thesis needs the strong view, but standard arguments support only the standard view.

Whither Pure Thinkers?

• So: prima facie and secunda facie, pure thinkers are possible.

2. What Could a Pure Thinker Think?

- Or: what sort of thing could a pure thinker think, if a pure thinker could think things?
- Cf. Strawson's *Individuals*: what sort of thing could someone with just auditory experience think?

Which Concepts?

- So far: a pure thinker could plausibly possess concepts from
 - logic (and, exists)
 - mathematics (plus, two)
 - thought (thinks, believes)
 - self (*I*)

Also

- It could also plausibly possess concepts from these domains:
 - time (now, before)
 - metaphysics (property, part, fundamental)
 - causal/nomic (cause, law, probability)
 - semantic (truth, reference)

Worldly Hypotheses

- A pure thinker could entertain thoughts about the world, not just its mind.
 - $\exists x: thinks(x) \& x \neq me$
 - $\exists y: cause(y, this thought)$
 - ∃ quantities q, r, s: q, r, s stand in such-andsuch nomic relations.

What Couldn't a Pure Thinker Think?

- sensory concepts: red, painful, loud
- spatial concepts (applied): tall, large
- perceptual demonstrates: this, that
- other singular concepts: Beyoncé, Montreal (though it might have descriptions that denote these individuals)
- practical demonstratives: this is how to

What Couldn't a Pure Thinker Think?

- Catherine Wilson on Descartes' immortal souls:
 - "If our minds endure after death, ... they will feel neither pain, nor pleasure, for they will no longer form a composite with our bodies. We will no longer see colours, touch objects, and hear sounds. We will not remember events of our past lives. We will be numb and inert. We humans will be almost nothing - at most capable of imageless thought and intellectual memory."



Pure Thinkers as Structuralist Thinkers

- Arguably, pure thinkers will be largely *structuralist* thinkers (about nonmental reality):
 - able to entertain structural hypotheses about the external world [like science according to structural realism]
 - akin to a red-green colorblind person who lacks the sensory concept *red* but still has a structural conception of redness.

3.What Can a Pure Thinker Know?

- Most obviously, a pure thinker can know
 - a priori truths e.g. about logic, mathematics, and other admissible topics
 - introspective truths about their thoughts
 - also: abductive truths about the world required to support these thoughts?

Upshot for Al

- Upshot: the unrestricted sense-thought thesis is false.
- An system without sensory grounding can
 - entertain a rich array of thoughts and build models of the world
 - refer (descriptively) to individuals
- though it will lack important sensory concepts and much empirical and practical knowledge

What About LLMs?

- LLMs aren't really pure thinkers: their textual I/O is arguably a textual sense with corresponding sensory concepts.
- LLMs can know about text inputs, and what explains them, and (if interpretable) then testimonial knowledge too?
- Structural+ thinkers, with expanded concepts and much more knowledge?

Part 2: The Poverty of Text

- LLMs seem to live in a world of text: ("In the beginning was the word...")
- Text alone cannot support understanding or meaning.

No Meaning from Form

 Bender and Koller: "the language modeling task, because it only uses form as training data, cannot in principle lead to learning of meaning"

Climbing towards NLU: On Meaning, Form, and Understanding in the Age of Data

Emily M. Bender University of Washington Department of Linguistics ebender@uw.edu

Alexander Koller Saarland University Dept. of Language Science and Technology koller@coli.uni-saarland.de

Ultra-Radical Interpretation

- Ultra-radical interpretation: determine what text from unknown speakers in an unknown language means, given just the text.
- Compare radical interpretation (Davidson, Lewis): determine what text ... means, given just utterances and behavior and environment (and internal processes)



Metaphysical vs Epistemological RI

- Ultra-radical interpretation looks much harder than RI.
- Construed as a metaphysical task (text constitutively determines meaning), it seems impossible: text doesn't entail meaning and hugely underdetermines it.
- However, construed as an epistemological task (text as a fallible guide to meaning), it's more feasible.

Ultra-Radical Translation

- Ultra-radical translation: translate text in an unknown language into one's own language, given just the text.
- Compare radical translation (Quine): translate utterances in an unknown language into one's own language, given just a speaker's behavior and their environment.



LLM Ultra-Radical Translation

- An epistemological version of ultra-radical translation has already been solved by LLMs, at least for pairs of human languages.
- Enough text from each language (unpaired, i.e. even without pairing translations in training) allows strong translation performance.
- Fallible, relies on contingent facts about the languages involved.

Epistemological Radical Interpretation

- Likewise, epistemological RI is easier than metaphysical RI.
- Form doesn't entail meaning, but hypotheses about meaning (and about the world) may explain form.
 - [cf. experience doesn't entail the external world, but external world explains experience.]
- There may be underdetermination, but a structural analysis of meaning (and of the world) will help minimize this.

NYC Subway System

- E.g. suppose we have a huge body of text Q&A about the NYC subway system along the following lines
 - "How do I get from 8th St/NYU to City Hall?"
 - "Three stops downtown on the N, R or W line".
- In a sophisticated language model, this data will naturally lead to structural hypotheses about the subway system and about the meaning of terms involved.





World and Meaning Models

- An LLM can use text data to infer (structural) world and meaning models.
 - World model: there exist entities e, properties Φ, ..., related in such-and-such structure
 - Meaning model: Term t denotes entity e in this structure. Predicate F denotes property Φ. Sentence F(t) is true iff e has Φ.

Implicit Representations of Meaning in Neural Language Models

Belinda Z. Li Maxwell Nye Jacob Andreas Massachusetts Institute of Technology {bzl,mnye,jda}@mit.edu

Where are the Facts Inside a Language Model?

Knowing differs from **saying**: uttering words by rote is different from knowing a fact, because *knowledge of a fact generalizes across contexts*. In this project, we show that factual knowledge within GPT **also corresponds to a localized computation that can be directly edited**. For example, we can make a small change to a small set of the weights of GPT-J to teach it the counterfactual "Eiffel Tower is located in the city of Rome." Rather than merely regurgitating the new sentence, it will generalize that specific counterfactual knowledge and apply it in very different linguistic contexts.





Part 2a: Stochastic Parrots

- This bears on a popular objection to LLM thinking (Bender, Gebru, et al):
 - they're just "stochastic parrots"
 - i.e. they are mere mimics or predictors
 - they only model text and don't model the world

On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?



Emily M. Bender^{1*}, Timnit Gebru^{2*}, Angelina McMillan-Major¹, Shmargaret Shmitchell³

¹ University of Washington ² Black in AI ³ The Aether *These authors contributed equally.

Training vs Processing

- It's true that LLM's are trained to minimize prediction error in string-matching. But their processing isn't just string-matching.
- Analogy: maximizing fitness during evolution can lead to novel processes post-evolution.
- Likewise: Minimizing string prediction error during training can lead to novel processes post-training (e.g. world-models).

Argument for LLM Thought

- Plausibly: an algorithm that truly minimized text prediction error (subject to constraints) would require deep models of the world and genuine thought and understanding.
- If so: sufficiently optimizing text prediction error in a language model should lead to world-models, thought, and understanding.

Upshot

- Ways out of the argument:
 - no (neural network) algorithms support thought
 - non-thinking algorithms outperform (or equiperform) thinking algorithms
 - the optimal algorithms can't be found through optimization methods
- Otherwise: thinking LLMs are possible, we just have to find them.

Part 3: LLMs and Knowledge-How

- Models in LLMs have many limitations: no robust sensory model of the world, no connection to embodied action.
- LLMs lack recognitional abilities and knowledge-how (except for non-embodied domains such as conversation).
Critiques of Al



LLMs and Know-How?

- Dreyfus's Heideggerian critique of AI and Adam's feminist critique both focus on the lack of knowledge-how in many AI systems.
- Are LLMs vulnerable to the Dreyfus/Adam critique? Yes and no.
- Yes: they lack skills and knowledge-how.
- No: they can easily be extended to embodied models with knowledge-how.

Knowing That and Knowing How

- Recent work: pure text LLMs can easily be adapted to recognition and action
 - Surprisingly little extra training required.
 - Translatable spaces for text, image, action
- The extensive text world of LLMs contains much of the background for skilled action?
- At least in machines, knowing that and knowing how are intimately linked?

Part 4: Limitations of Current LLMs

• If possible LLMs can think (in a limited way), what about current LLMs?

Limitations of Current LLMs

- (Lack of senses and embodiment.)
- Many bad glitches in reasoning.
- Lack of recurrence and genuine memory.
- Lack of internally-driven thought.
- Lack of stable goals and unified agency

Can Current LLMs Think?

- These may be good reasons to deny that current LLMs can think, pending separate analysis (both of LLMs and of thinking).
- Alternatively, current LLMs might be borderline cases of thinkers, as e.g. fish might be.
- Many of these limitations will be overcome in the next decade or two.

Conclusions

- There's a prima facie case that in principle (if not yet in practice), LLMs can think and understand.
- Arguments from sensory grounding don't establish that LLMs can't think or understand though they impose substantial restrictions.
- These restrictions might be overcome by extended LLMs with senses/embodiment.
- It's a serious possibility that AI systems will have genuine thought and understanding in the coming years and decades.