DENKWERKSTATT BERLIN

Anna Strasser

What smart AIs can do without consciousness

Main question

Is having consciousness a necessary condition to explain the performance of AI systems?

Generative AI, in particular LLMs, appear to be able to solve all kinds of tasks for which humans need a range of socio-cognitive abilities, such as reasoning, planning, and understanding.

IN HUMANS, SUCH ABILITIES SEEM TO BE NECESSARILY ASSOCIATED WITH CONSCIOUSNESS.

Just because humans need consciousness, machines may not need consciousness to do something comparable

We might not need to ascribe consciousness to the novel artificial systems developed with generative AI (GenAI) in order to explain their performances.

Slides are downloadable at

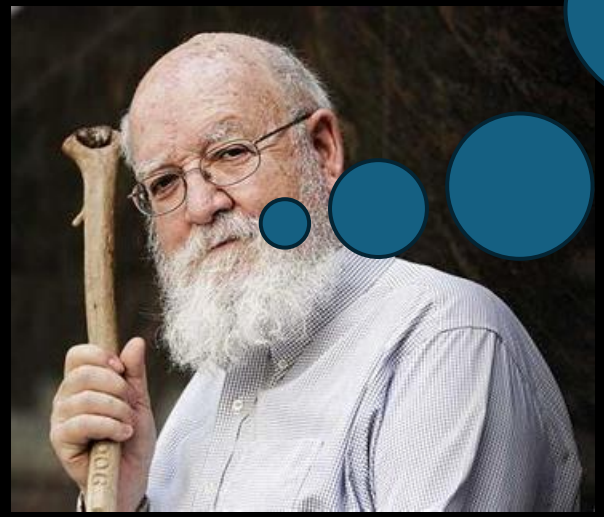
Overview

1. • Gradualism & Inbetweenism
2. • Why AI consciousness might not be the first question to be discussed
3. • Towards a multidimensional spectrum of agency and intelligence
4. • Multiple realizations situated in a multidimensional spectrum
5. • Still in search of a justified ascription

Gradualism

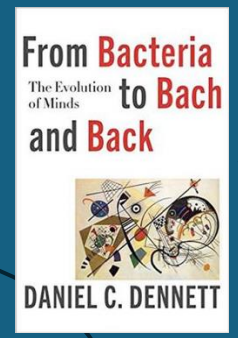
The performances of GenAI tools challenge philosophy to come up with an appropriate characterization of their properties and abilities.

Like Daniel Dennett, I aim to argue for a gradualist approach towards comprehension and claim that *"comprehension comes in degrees"*



I recommend we discard this way of thinking. This well-nigh magical concept of comprehension has no utility, no application in the real world. But the distinction between comprehension and incomprehension is still important, and we can salvage it by the well-tested Darwinian perspective of gradualism: ***comprehension comes in degrees****. At one extreme we have the bacterium's sorta comprehension of the quorum-sensing signals it responds to (Miller and Bassler 2001) and the* ***computer's sorta comprehension*** *of the "ADD" instruction. At the other extreme we have Jane Austen's comprehension of the interplay of personal and social forces in the emotional states of people and Einstein's comprehension of relativity.*

(Dennett, 2018, 95)

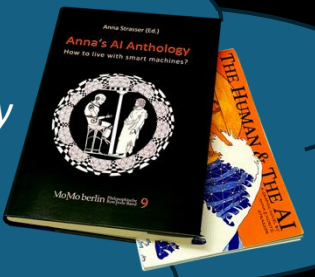


This is a strong motivation to develop new adequate notions with which we can describe *sorta comprehension* when addressing the competence of artificial systems.

Conceptual problem

[…] it is neither quite right to say that our interactions with LLMs are properly asocial (just tool-use or self-talk) nor quite right to say that our interactions with LLMs are properly social. Neither standard philosophical theorizing nor dichotomous ordinary concepts enable us to think well about these in-between phenomena.

(Strasser & Schwitzgebel, 2024, 197)



Characterizations of the properties & abilities of novel artificial systems pose a conceptual problem, because we do not have the right notions to describe them.

My idea is that we should be careful to not end up as victims of build-in limitations of our contemporary conceptual frameworks that lead in my view either to overattributing or underattributing properties and abilities to AI systems.

A multidimensional spectrum of social interactions

CAN BE CONCEPTUALIZED WITH THE HELP OF A DISJUNCTIVE CONCEPTUAL FRAMEWORK

TOOL USE

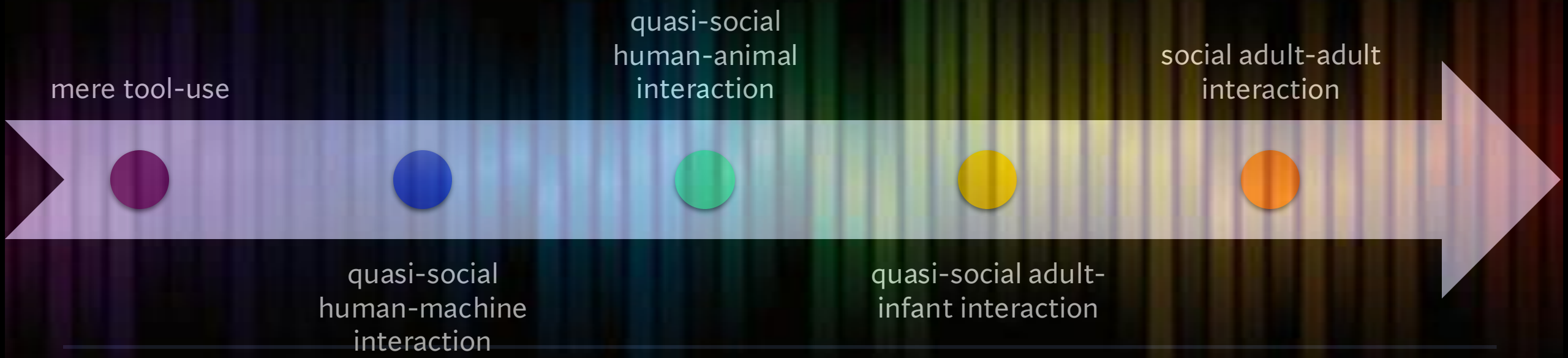
INBETWEEN PHENOMENA

SOCIAL INTERACTION

SINGLE-SIDED SOCIALITY

QUASI-SOCIAL ASYMMETRIC INTERACTIONS

FULL-BLOWN, INTELLECTUALLY DEMANDING, COOPERATIVE SOCIAL INTERACTION



Inbetweenism

→ INSTANCES IN ASSUMED MULTIDIMENSIONAL SPECTRA MAY QUESTION THE NECESSITY OF CONSCIOUSNESS

FAMILY RESEMBLANCE

- all the instances in a multidimensional spectrum stand in a relation of family resemblance
- they do not need to fulfill the very same list of conditions

Wittgenstein, Ludwig. 2009. *Philosophical investigations*.

SPECTRUM OF AGENCY

- minimal/quasi/sorta cases
  - questioning the necessity of consciousness
- full-fledged cases as we find them in humans
  - fully developed agency that is closely connected to consciousness

A DISJUNCTIVE CONCEPTUAL FRAMEWORK

- does not require a whole package of conditions that necessarily co-occur
- allows for various combinations of conditions that can capture the diversity of phenomena

SPECTRUM OF INTELLIGENCE

- minimal/quasi/sorta cases
  - imagining spectra of all the socio-cognitive abilities that contribute to intelligence
- full-fledged form of human intelligence

Approaching INBETWEEN phenomena as instances in a spectrum allows us to grasp the similarities and differences between artificial and human actors in a more nuanced way.

Do agency, intelligence, and consciousness necessarily cooccur?

AVOID CONFUSIONS BETWEEN THE ATTRIBUTION OF AGENCY, INTELLIGENCE, AND CONSCIOUSNESS

- agency & intelligence may not presuppose consciousness
  - even though up to now the full-fledged forms of agency/intelligence are found in humans with consciousness

➢ cooccurrence of agency & intelligence with consciousness is perhaps a peculiarity of humans

ARTIFICIAL SYSTEMS ARE DIFFERENT

- do not solve tasks in the very same way as humans
- have significantly more training data at their disposal than a single human can process in their lifetime
- processing speed is much faster than that of humans
- do not have a history of growing up as social beings in a community
- most of them lack any form of embodiment

→ striking differences in their information processing

ARTIFICIAL SYSTEMS ARE SIMILAR

- have forms of agency and intelligence
- can do a lot what humans can do

→ striking similarities in their outputs

Instances of agency & intelligence should be imagined as being located somewhere in a multidimensional spectrum that allows for distinct combinations of conditions

Unknowable AI consciousness?

no agreement on what consciousness is

- higher-order theory (Rosenthal, 2005)
- global workspace theories (Mashour et al., 2020)
- integrated information theory (Tononi et al., 2016)
- and many more (Seth & Bayne, 2022)

no agreement on how to measure it

- variety of tests for consciousness (Bayne et al., 2024; Ferrante et al., 2025; Schneider, 2019)

I think that it does not look as if we would come to an agreement with respect to AI consciousness any time soon

Forthcoming book by Eric Schwitzgebel: AI and Consciousness

In this book, I aim to convince you that the experts do not know, and you do not know, and society collectively does not and will not know, and all is fog.

Two extreme positions

Hardcore instrumentalist view

no agency
no intelligence
no consciousness

- excluding the possibility that any artificial system could have a social status in an HMI

In expection of AGI view

agency
intelligence
consciousness

- whole demanding package of conditions that we require from humans can, in principle, also be fulfilled by sophisticated machines

Things don't dichotomize

PHILOSOPHY POSES TOO DEMANDING CONDITIONS

FULL-FLEDGED AGENCY BY DONALD DAVIDSON

INTELLECTUALIST CONCEPTIONS

- ❖ philosophers tend to describe ideal cases that are rarely found in everyday life
- ❖ children, non-human animals, and robots (artificial agents) tend to fall through the conceptual net

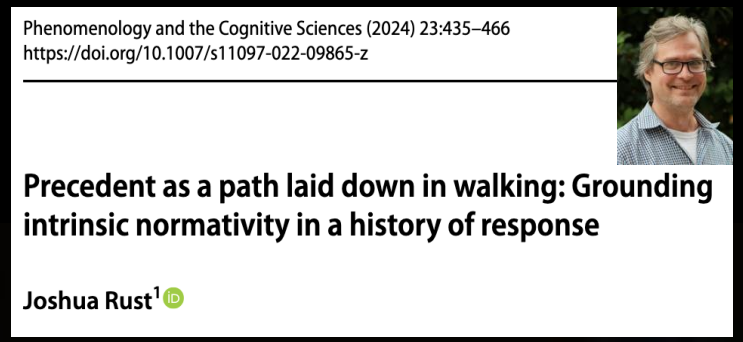
GRADUALIST APPROACHES & MINIMAL NOTIONS

- ➢ agency seems to be something that does not emerge in an instant

be that developmentally in humans, phylogenetically in animal evolution, or technologically in the design of AI systems

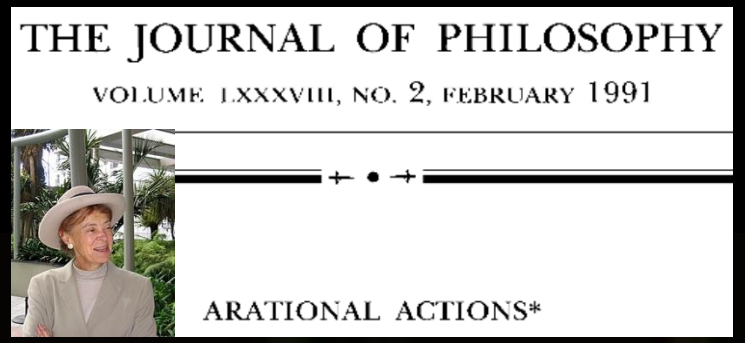
Artist: Lorin Strasser

Various kinds of agency


Phenomenology and the Cognitive Sciences (2024) 23:435–466
https://doi.org/10.1007/s11097-022-09865-z

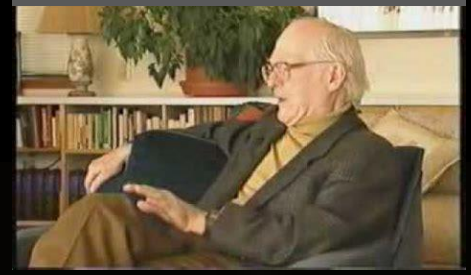
Precedent as a path laid down in walking: Grounding intrinsic normativity in a history of response

Joshua Rust[1]


THE JOURNAL OF PHILOSOPHY
VOLUME LXXXVIII, NO. 2, FEBRUARY 1991

ARATIONAL ACTIONS*

FULL-FLEDGED AGENCY BY DONALD DAVIDSON



quasi agency

joint agency

precedential agency

arational agency

agency of children and non-human animals

full-fledged agency

Including non-living entities?

MINIMAL APPROACHES

Stephen Butterfill & Ian Apperly (2013): minimal mindreading | John Michael et al. (2016): minimal sense of Commitment | Elisabeth Pacherie (2013): shared intention lite | Dominik Perler & Markus Wild (2022): simple minds

UTILIZING MINIMAL APPROACHES TO DESCRIBE VARIOUS SETS OF CONDITIONS

characteristic feature :

- questioning the necessity of some conditions
- allow for a less strong manifestation
- connect empirical findings and our common sense with theoretical work in philosophy

Quasi actions

Is it necessary to have the ability to set goals in order to be an agent?

AGENTS WITH A MINIMAL FORM OF AGENCY

Cognitive systems with a flexible coupling between input and output

- implies
  - learning abilities to adapt to environmental changes in a dynamic world and acquire knowledge in relation to an action goal
  - cognitive abilities of evaluation, planning, anticipation, and trial and error

AGENCY & INTELLIGENCE COOCCUR

→ To prove themselves capable of acting in our world, they need
- the ability to take in relevant information and represent it in a world model
- effectors that can cause changes in the environment
- to be autonomous to a certain degree

NOTION OF ACTION IS CLOSELY INTERWOVEN WITH THE CONCEPT OF THE ENVIRONMENT

Our world is a dynamic, complex environment.

- Simple *block world environments* do not represent dynamic, complex environments.
  - As impressive as chess or go-playing systems are, they operate within an environment that is fixed by rules. Therefore, they cannot demonstrate their agency through adapting to a dynamic environment.
- In HMIs with GenAI technology products, e.g., in the interactions between humans and LLMs, artificial systems can prove themselves capable of acting in a dynamically developing environment – our world of language games.

Without cognitive abilities, no entity can prove as an agent, at least not in a dynamic, complex environment.

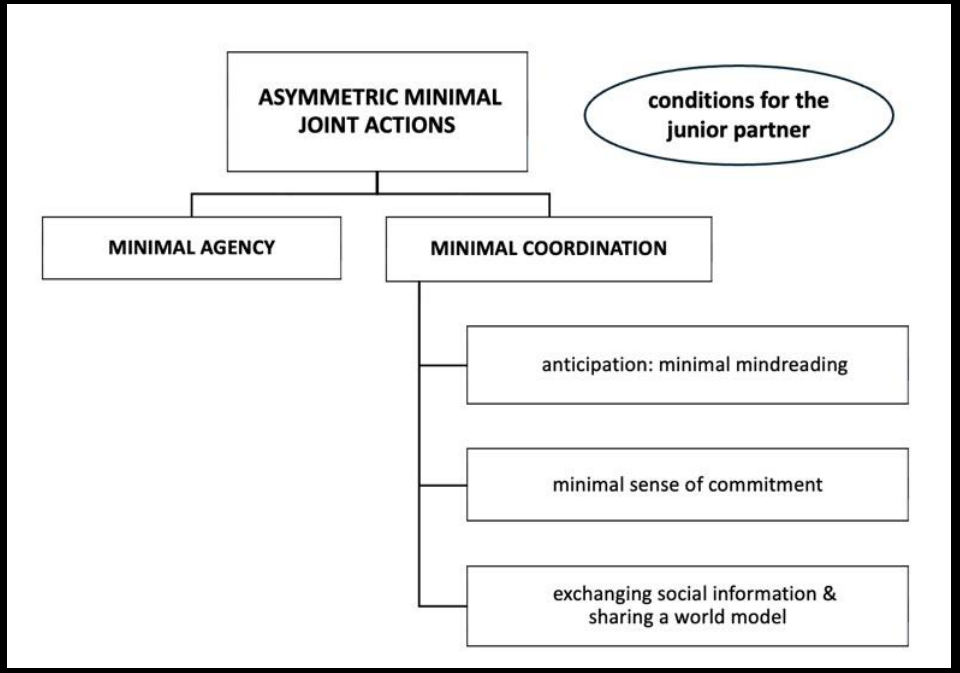
Intelligence needed by quasi-social interaction partners

For quasi-social interactants with quasi joint agency

- instrumental rationality is sufficient
  - autonomously develop sub-goals that contribute to the success of an HMI
- no reflective rationality needed
  - only full-fledged agents need to be able to generate goals

For asymmetric joint actions

- minimal agency
- minimal abilities to coordinate with the human partner
  - issues like anticipating the behavior of the partner, allowing for some sort of commitment, contributing to an exchange of social information, and sharing a world model



No necessity to consider them sentient beings with feelings and desires.

Interim balance

ASSUMING MULTIDIMENSIONAL SPECTRA
in which instances stand in a relation of family resemblance

Humans cannot participate in social interactions without being conscious

attribute full-fledged agency, intelligence and consciousness

AI systems can qualify as quasi-social interactants without being conscious

attribute a minimal form of agency and specific cognitive abilities to AI systems without being tempted to see them as conscious beings

- We feel strongly invited to attribute the whole package of agency, intelligence, and consciousness to them as we do if we play language games with our fellow humans.
- But I think that we should aim to consider differences and similarities if we interact with AI systems and be open to the idea that there are asymmetric interactions in which participating agents fulfill distinct sets of conditions.

IF NOT ALL CONDITIONS that characterize full-fledged forms HAVE TO CO-OCCUR if we describe other instances in a multidimensional spectrum THEN we may question whether consciousness is for all instances a necessary condition.

- Still in search of a justified ascription

A huge remaining problem

Arguing for a disjunctive conceptual framework that enables us to characterize a diversity of instances in a multidimensional spectrum is one thing.

To make use of it, one needs, of course, an idea of how one can argue for justified ascriptions.

- Any application of a conceptual framework raises empirical questions, namely, questions about the extent to which certain conditions (abilities and properties) are fulfilled by the entity in question.

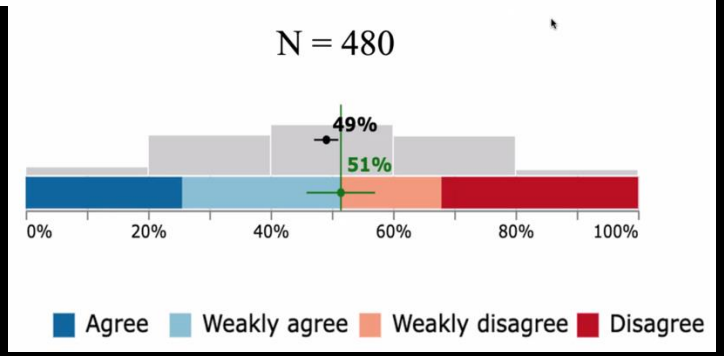
SHOULD WE ASK THE CREATORS OF ARTIFICIAL SYSTEMS?

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.



Being able to construct smart AI systems does not necessarily come along with an understanding of their properties and abilities that make their performance possible.

CONTROVERSIAL DEBATES

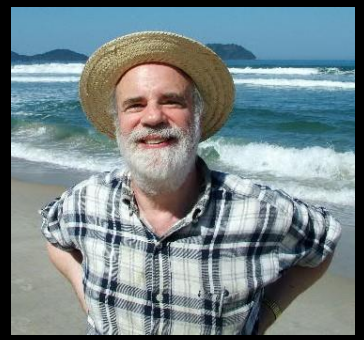
- no agreement on the question whether the statistical processing of training data leads to a multiple realization of socio-cognitive abilities

Routes not to be taken

OBSERVING INPUT-OUTPUT PATTERNS IS NOT SUFFICIENT

Even if the output is similar to the output humans deliver, we cannot be sure of how this output is achieved.

rule-following paradox
(Wittgenstein / Kripke)



Are LLMs quadding or adding?

LLMs that perform well in a benchmark might still follow different rules than humans would follow to solve such tasks

→ **JUST OBSERVING INPUT-OUTPUT PATTERNS IS NOT SUFFICIENT**

benchmarks come with critical issues

- data contamination
- robustness of the results
- problems with flawed benchmarks

machine might make use of

- memorization
- shortcut learning
- subtle statistical associations

WE SHOULD BE CRITICAL OF WHETHER BENCHMARKS ACTUALLY MEASURE WHAT THEY CLAIM TO MEASURE

Beyond input-output patterns

WE NEED TO INVESTIGATE THE PROCESS BY WHICH THE PERFORMANCE IS ACHIEVED

mathematical descriptions do not lead to useful insights into whether the performance is due to the possession of socio-cognitive ability

- no human-intelligible descriptions by which one could decide whether socio-cognitive abilities have emerged

detailed description of the human brain at the molecular and cellular levels

mathematical descriptions

of a huge composite function consisting of a complex sequence of linear and nonlinear transformations across many layers

taking a physical stance towards human beings does not exclude the possibility that we are justified to take an intentional stance towards them

being able to give a mathematical description of neural nets does not yet exclude that they might possess socio-cognitive abilities

Contra arguments stating that because LLMs' operations can be described by a mathematical description that refers to statistical calculations, linear algebra operations, or next-token predictions, those descriptions are also all we could ever ascribe to them.

Interpretability techniques

AIM TO UNCOVER THE CAUSAL MECHANISMS UNDERLYING LLMs' PERFORMANCE AT A HIGHER LEVEL

investigating the inner structure of neural networks by asking whether LLMs

- represent information,
- operate on representations,
- have activation patterns that realize socio-cognitive abilities

A very accessible presentation of the details of such approaches can be found in *A Philosophical Introduction to Language Models* (Millière & Buckner, 2024b, 2024a)

probing

- exploring what is encoded in a neural network.

➢ statements that certain information is likely to be represented in their activation pattern

BUT does not yet provide information as to whether these representations are used when the model solves a task.

attribution methods

- explore which parts of the input data (the prompts provided by the human interaction partner) a model relies on most for their outputs

causal intervention methods

- determine the causal role played by a representation in the processing of a model
  - models are changed in various ways, and it is examined whether the intervention changes the predictions (the outputs) of the model in a systematic way
    → hypotheses regarding the processing are tested, e.g., whether a model performs a systematic calculation to solve the task or whether a system has something like a world mode

TWO DIFFICULTIES

- techniques are mostly practiced with toy models → wait until they are applied to large language models
- rely on operationalizable theories of all the abilities we want to ascribe to LLMs

A plea for cross-disciplinary approaches

At this point, one could despair and say that we are staring into an abyss

Little hope that we will ever be able to build conceptual bridges

ATTRIBUTION STRATEGIES ARE ALSO SOCIAL PRACTICES

- ➢ do not only depend on scientific progress
  - but are also shaped by developments in our common sense

- → aim for cross-disciplinary collaborations
  - that involve not only philosophy and computer science
  - but also psychology, sociology, and legal sciences

Both our common sense and scientific progress
have the potential
to shape the meaning of the words
we use to describe
socio-cognitive abilities & agentive properties

This uncertainty regarding the justified attribution of properties and capabilities motivates cross-disciplinary collaborations that might lead to a commonly agreed-on practice.

Conclusion

Do all ascriptions of agency and intelligence necessarily presuppose the ascription of consciousness?

IF we utilize the strategy of how well-established gradual approaches

- expanded the notions of agency and various socio-cognitive abilities in a way that makes them applicable to non-human animals and children by questioning conditions that are considered necessary by standard intellectualist notions, and allowing for varying degrees of the manifestation of conditions

➢ **THEN** we may develop notions that are applicable to artificial systems without requiring consciousness as a necessary condition.

➢ **MULTIDIMENSIONAL SPECTRUM IN WHICH A DIVERSITY OF INSTANCES CAN BE LOCATED**

➢ BUT THIS DOES NOT solve the perhaps much harder problem of when we can justifiably attribute to AI systems the properties and abilities

- Expertise in computer science alone is not sufficient to arrive at a recognized practice in the attribution of characteristics and abilities

➢ motivate us to explore how other disciplines like psychology, sociology, and legal sciences may help us elaborate on further factors that can play a role in agreeing on an attribution practice

Takeaway

If you have agency & intelligence and you are an artificial system, you might not get consciousness for free.