THE PHILOSOPHY OF DEEP LEARNING

NEW YORK UNIVERSITY | MARCH 24-26, 2023

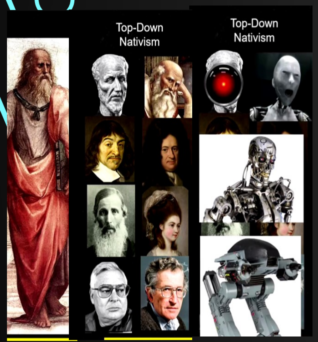
ORGANIZED BY NED BLOCK (NYU), DAVID CHALMERS (NYU), RAPHAËL MILLIÈRE (COLUMBIA UNIVERSITY)

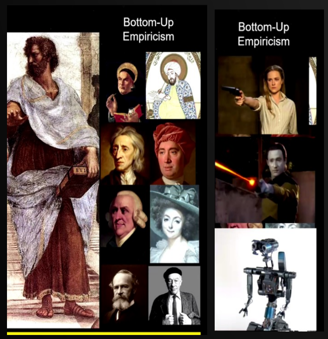
Memory slices by Anna Strasser

DISCLAIMER: JUST MEMORIES – AIMING FOR CORRESPONDENCE WITH REALITY BUT CANNOT GUARANTEE IT.

Cameron Buckner (Houston)

Moderate empiricism & machine learning





INNATE vs LEARNED | NATURE vs NURTURE

innate

manually program core knowledge

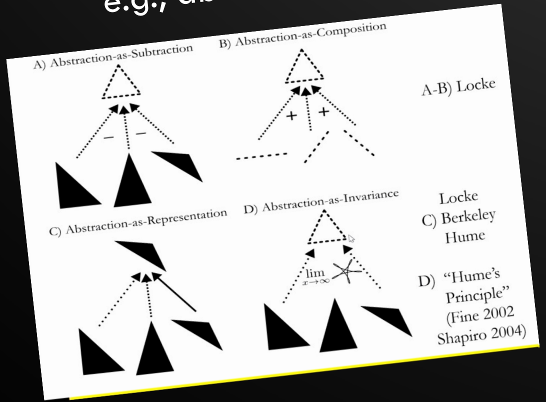
consequences for engineering methodology

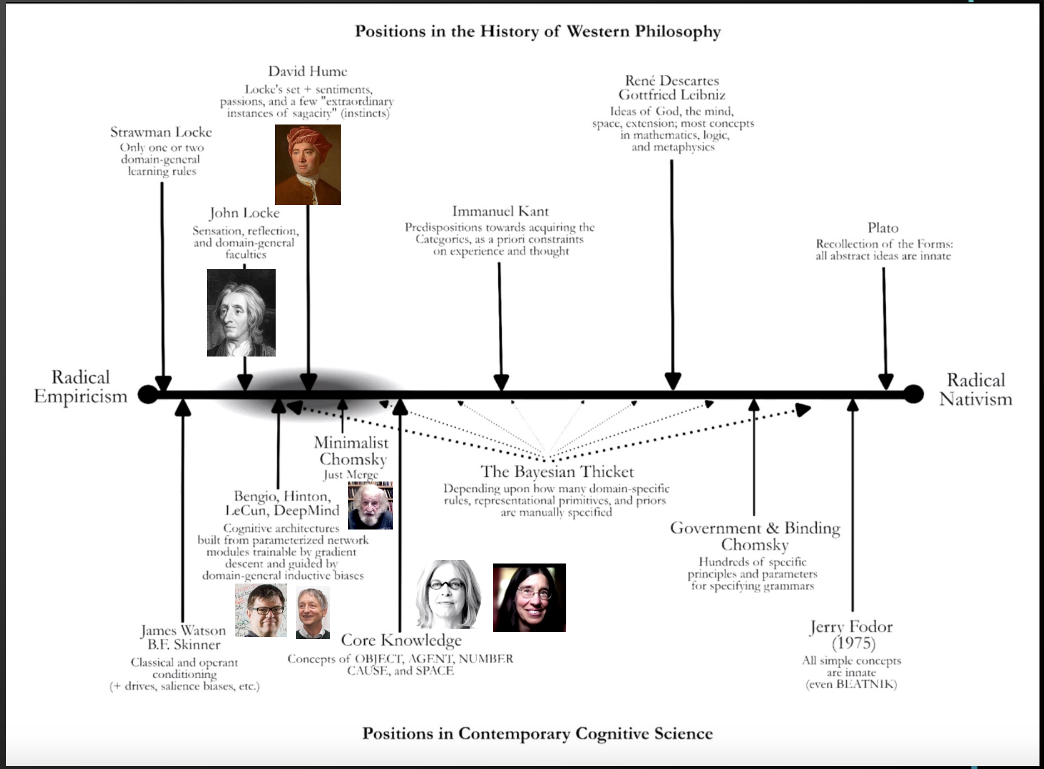
ML might mine gems from the history of philosophy (general cognitive architecture & rational decision-making

derived from sensory experience

enable them to learn* the domain-specific abstractions themselves

ORIGIN OF ABSTRACT KNOWLEDGE
e.g., abstract triangle





READ UPCOMING BOOK

Cameron J. Buckner (2023). Deeply Rational Machines. What the History of Philosophy Can Teach Us about the Future of Artificial Intelligence. Oxford University Press.

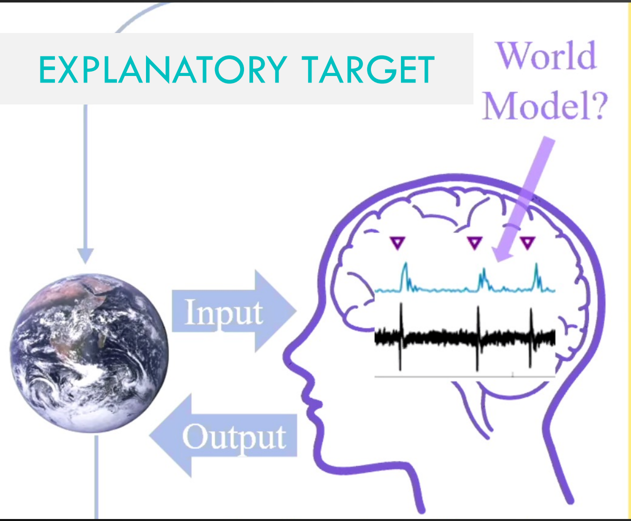
Rosa Cao (Stanford)

Are apparently successful DNN models also truly explanatory?

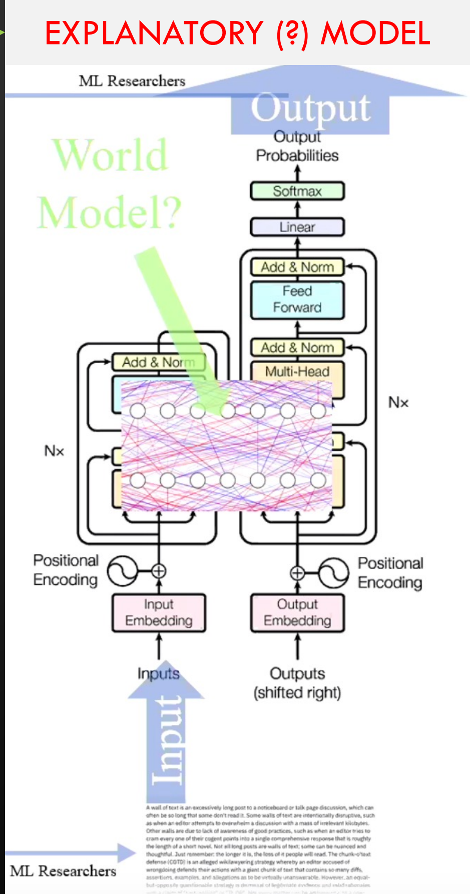
Do models have understanding? Do their words have meaning? Are they (relevantly) like us? Do they have representations with the same functional role (e.g., inner models structuring behavior)?

INSTEAD OF

- "Stochastic parrots"
- "Mere next-word prediction"
- "Capturing surface patterns"
- "Curve-fitting"



- What aspects of your target does the model capture?
- To what degree?
- Under what assumption?
- How robust is your model?
- How well does it generalize?
- How efficient is it?



Representational pragmatism

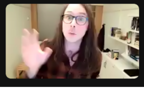
patterns of activity

- should be causally involved in behavior
- must be manipulable at the representational level
- ➢ ascriptions are relative to a probe (& explanatory purpose)

Symposium: Representation in Deep Learning Systems

Fintan Mallory (Oslo)

Teleosemantics for Neural Word Embeddings



In conclusion…. these are the same thing

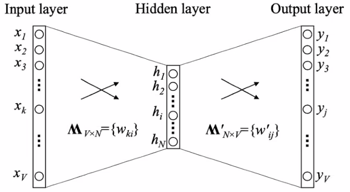
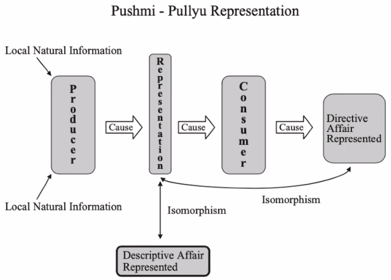


Figure 1: A simple CBOW model with only one word in the context



Rong, X., 2014. word2vec parameter learning explained. arXiv preprint arXiv:1411.2738. (slight modification)

Millikan, R. G. (2005). *Language: A biological model*. Oxford University Press

zoom.us

Jacqueline Harding (Stanford)

Summary

To assess whether component h represents a property Z:

- **(Information)** Train a successful probe $g_Z : h(D) \to \mathcal{P}(Z)$.
- **(Use)** Apply an `ablate` intervention to $h(s)$ for $s \in D$. See if system's performance degrades.
- **(Misrepresentation)** Apply a `correct` intervention to activation $h(s)$ for $s \in D$. See if system's performance improves.

Anders Søgaard (Copenhagen)

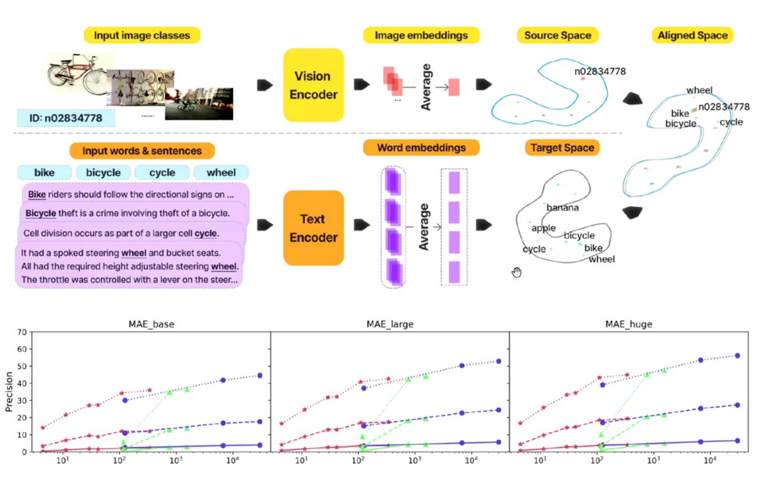
A response to Bender & Koller (2020). Climbing towards NLU.

UNSUPERVISED MACHINE TRANSLATION

1. vocabulary alignment using point set registration algorithms
2. co-reference of 'line' and 'linea'
3. translate

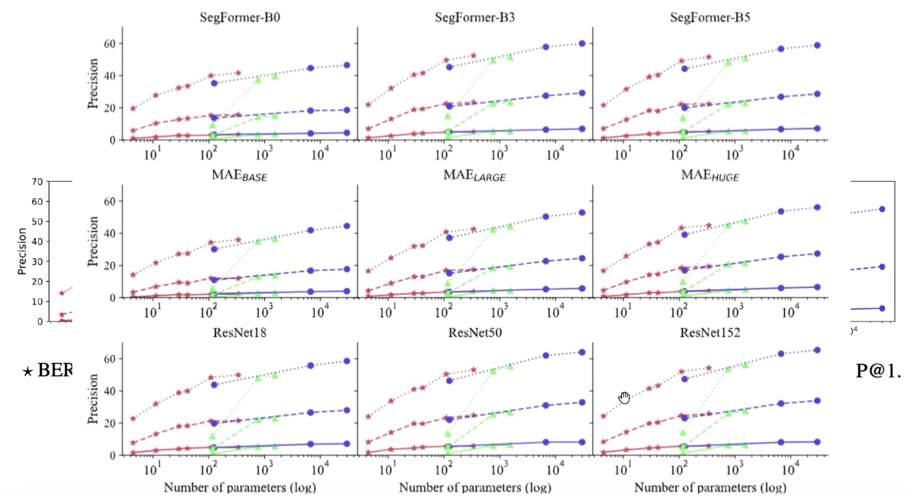
BUT works only if spaces are very

Key idea: If LM and CV models were aligned in the same way, we could translate and do VQA.



⋆ BERT models, ▲ GPT2 models, ● OPT models; Dotted line: P@100, dashed line: P@10, solid line: P@1.

SUPERVISED



Models	Polysemy	Pairs	SegFormer-B5 P@100	MAE$_{HUGE}$ P@100	ResNet152 P@100	Dispersion	SegFormer-B5 P@100	MAE$_{HUGE}$ P@100	ResNet152 P@100
BERT$_L$	1	100.8	**58.5**	**60.2**	**61.7**	low	**60.4**	**57.1**	**61.7**
	2-3	178.4	46.4	47.4	49.3	medium	48.3	49.5	52.5
	4+	319.6	37.3	36.5	39.7	high	28.6	28.4	30.7
GPT2$_{XL}$	1	100.8	**54.6**	**55.5**	**58.5**	low	43.2	47.6	49.5
	2-3	178.4	52.6	52.7	54.4	medium	**49.1**	**52.2**	**54.4**
	4+	319.6	37.7	40.1	42.5	high	41.1	42.3	45.2
OPT$_{30B}$	1	100.8	**64.3**	**65.17**	**68.8**	low	**60.4**	**60.0**	**68.0**
	2-3	178.4	56.3	56.9	59.2	medium	56.4	59.9	62.4
	4+	319.6	39.1	41.5	44.7	high	38.6	46.8	44.9

Is this knowledge?

Control experiment: Could it be that LMs and VMs are contaminated by inductive bias or ImageNet artefacts? To check, we ran similar experiments mapping BigGraph embeddings into LM vector spaces - obtaining very similar results. This suggests the convergence is not explained by contamination or ImageNet artefacts.

Language model	P@1	P@10	P@100
{BERT-Tiny}	1.05263	10.52632	35.26316
{BERT-Mini}	2.10526	11.05263	38.94737
{BERt-Small}	2.63158	14.73684	41.57895
{BERt-Medium}	1.57895	13.15789	46.84211
{BERT-Base}	0.0	17.89474	53.68421
{BERT-Large}	2.10526	19.47368	55.05263

Søgaard (2023): '**Grounding the Vector Space of an Octopus**'. Minds and Machines.
Li et al. (2023): '**Implications of the Convergence of Language and Vision Model Geometries**'. ArXiv.

Tony Chen, Mitchell Ostrow, Hokyung Sung, Cedegao Zhang

Do deep neural networks have concepts?

EMPIRICAL TEST FORMAL CHARACTERIZATION OF CONCEPTS

HOW SOME FEATURES MIGHT BE PARTIALLY OPERATIONALIZED

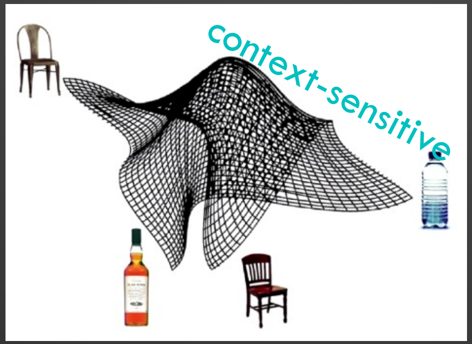
Barsalou 2020

The conceptual system should allow for sampling referents or tokens of any concept.

generative

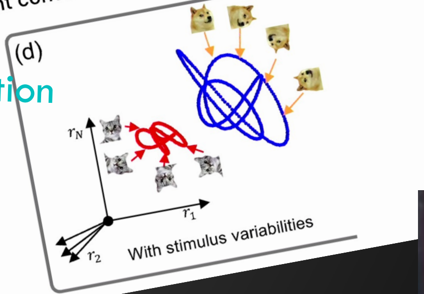
Manifold view: there exists a probability distribution over the concept manifold that allows the system to sample from it.

De Martino et al. 2023



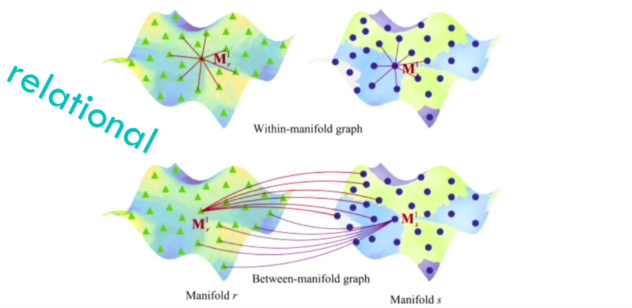
Hofstadter & Sander 2013, Chung & Aboutt 2021, Odouard & Mitchell 2022

The concept manifold should preserve invariance with respect to irrelevant contextual inputs or transformations.



Shi 2020

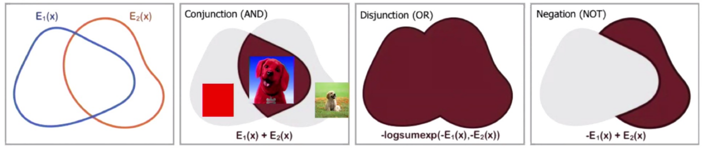
The relations between concepts is captured by the geometry of the overall latent space, which includes multiple concept manifolds.



Lake & Baroni 2018, Hupkes et al. 2019, Lewis et al. 2022

The compositional operators **and**, **or**, and **not** correspond to manifold intersection, union, and complement

compositional



OTHER FEATURES

- discriminability
- intentional, consistent, causal structure

Some of these properties are incredibly important and of philosophical and psychological interest, but it is not clear how they might be formalized.

Panel:

What Can Deep Learning Do for Cognitive Science and Vice Versa?

Speakers:

- Ishita Dasgupta (DeepMind)
- Niko Kriegeskorte (Columbia)
- Tal Linzen (NYU / Google AI)
- Robert Long (Center for AI Safety)
- Ida Momennejad (Microsoft Research)

Ishita Dasgupta (DeepMind)

Wie et al. (2022). Chain of thought prompting elicits reasoning in large language models.

Homo economicus
Perception as Baysian inference

SPECTRUM OF RESPONSES

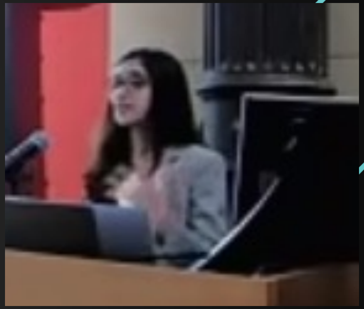
LLMs are...

... capable of reasoning ⟷ ... using simple heuristics

Humans are...

... capable of reasoning ⟷ ... using simple heuristics

Valmeekam et al. (2022). Large Language Models Still Can't Plan.

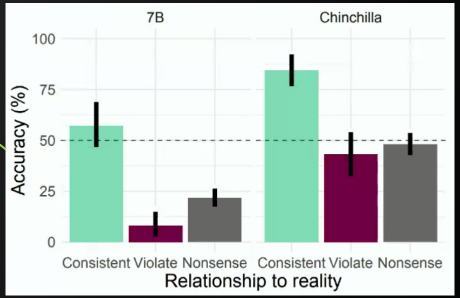


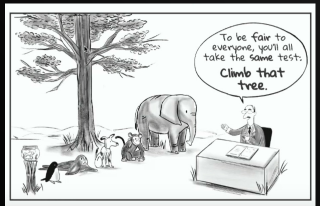
Predictibility
Thinking slow / thinking fast

humans are better at reasoning in familiar social settings (Wason task)

- LLMs have prior expectations over language; that's their point.
- LLM expectations often **reflect** human beliefs & knowledge.

Will LMs show the same content effects on reasoning as humans?



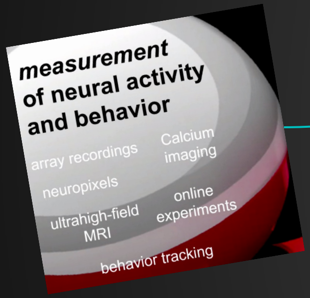


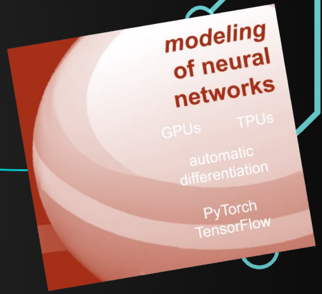
What can we learn from this?

- These effects can emerge from a **monolithic** model, trained on a **simple** task objective – without explicit dual systems or social reasoning mechanisms. How this emerges in LMs is worth understanding, to understand it in humans.
- Developing new levels of analysis:
  similar "behavior" < similar "representations" < similar "learning"
- Cognitive science has vocabulary and empirical methodology to yield insights for current AI – or at least its applications.
- A new comparative psychology?

Niko Kriegeskorte (Columbia)

DISRUPTED BY TWO REVOLUTIONS



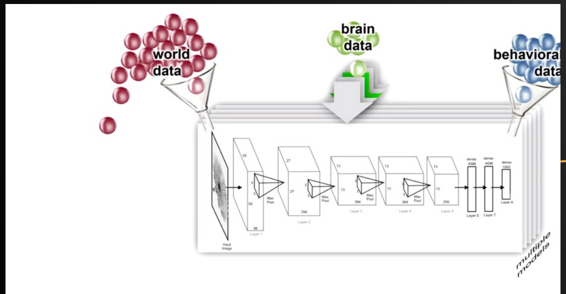


just data fitting!

Neural network models as **mechanistic explanations**

too simple!
(not faithful to biology)

too complex!
(not intuitively explainable)

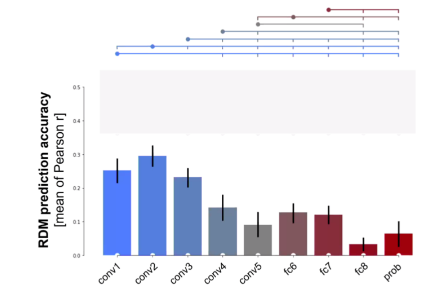


Kriegeskorte 2015, Kriegeskore & Douglas 2018

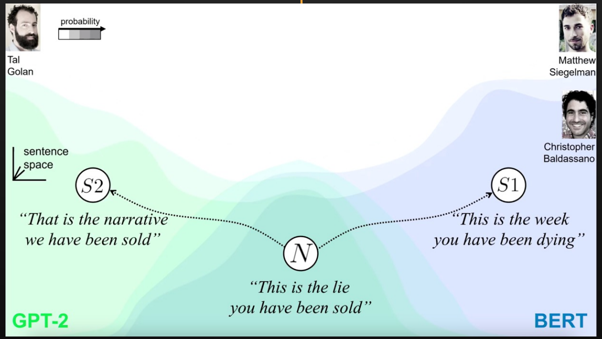
Conclusions

1. **Neural network models promise *mechanistic explanations*** of brain-information processing, but theoretical progress requires new methodology for comparing high-parametric neural network models.
2. ***Model-comparative inference*** that generalizes across experimental conditions and subjects enables progress toward better models and theories.
   Schütt et al. pp2021
3. **Optimized experiments using *controversial stimuli*** provide severe tests of out-of-distribution generalization for different deep net models.
   Golan et al. 2020

Model comparison



Schütt et al. pp2021
RSA3 open-source Python Toolbox in collaboration with the labs of Diedrichsen, Mur, and Charest.



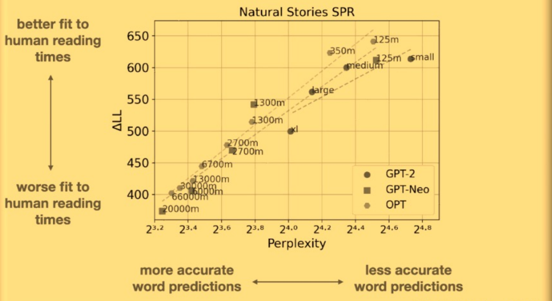
Tal Linzen (NYU / Google AI)

What, if anything, can LLMs teach us about human language acquisition?

Modern neural networks are stronger learners than the cognitive models we had in the past—we can just unleash them on a corpus, without simplifying or annotating it

- → Which assumptions lead to the successful acquisition of linguistic generalization?
- → Do we need Universal Grammar?
- → Do we need perceptual grounding?
- → What representations emerge to support the network's behavior?

Predictions from larger language models are increasingly non-human-like



(Oh & Schuler, 2022)

But we need to be able to control the assumptions: commercial "large" language models are increasingly unhelpful here

a useful infrastructure

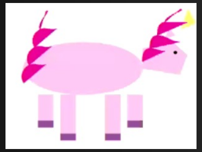
IF MODELS ARE TRAINED ON HUMAN-APPROPRIATE DATA

- e.g. resource-limited in human-like ways
- not the ones corporations find attractive

EXPERIMENTS WITH COMMERCIAL LLMS ARE NOT RELEVANT

Robert Long (Center for AI Safety)

Why cognitive science is not helful for AI



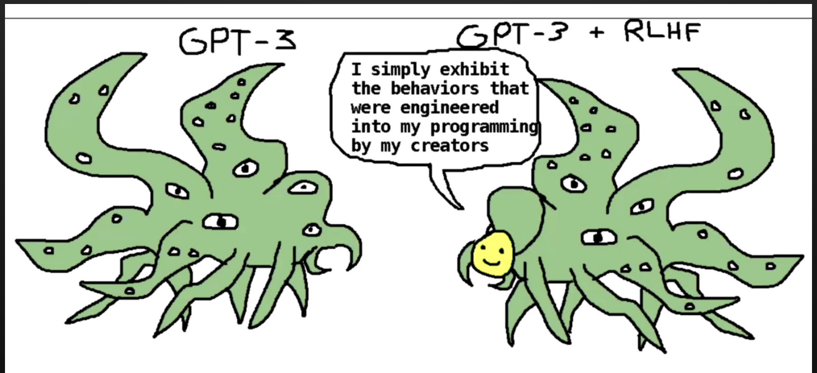
VALUABLE INSIGHTS ABOUT THE COMPUTATIONAL BASIS OF HUMAN (AND ANIMAL) INTELLIGENCE

- reverse engineering
- transferrable insights from neuroscience, philosophy, etc.
- cognitive science: plausible & appealing but false in practice
- AI systems don't need those solutions … especially not at scale

- There are **principled reasons** to expect it to be false
  - 1) We are not good at cognitive science
  - 2) AI systems have little use for built-in human-like solutions, *especially at scale*
- (and this makes me sad)

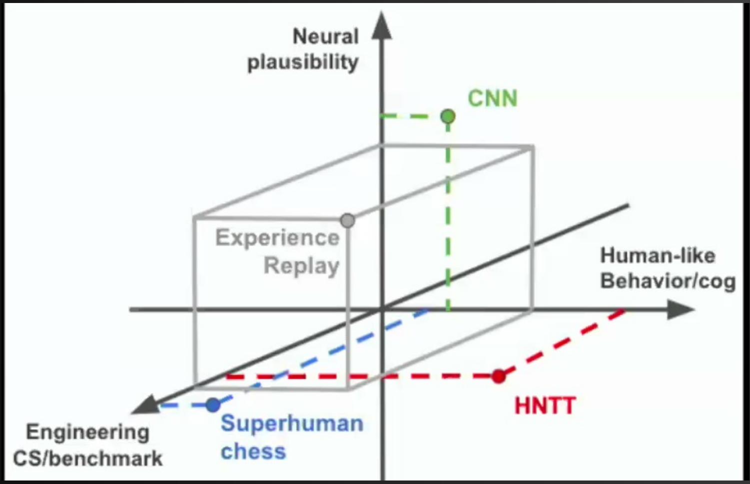
1) The computational basis of human intelligence is far **more complex** than our theories in cognitive science have captured
   - Leads to brittle 'solutions' when applied

2) Human-like solutions are **not optimal** for AI systems
   - Human-like solutions are optimal given human:
     - **Computational capacity**
     - **Data**
     - **Timescale** of learning
   - Imposing human-like constraints - like all constraints - predictably becomes unhelpful with scale (Sutton's "Bitter Lesson")



Ida Momennejad (Microsoft Research)

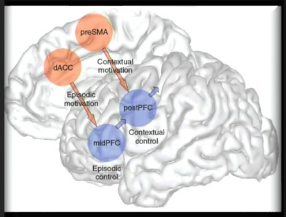
LLMs need a [dumb] PFC



Momennejad, I. (2023). A rubric for human-like agents and NeuroAI. *Philosophical Transactions of the Royal Society B*, *378*(1869), 20210446.

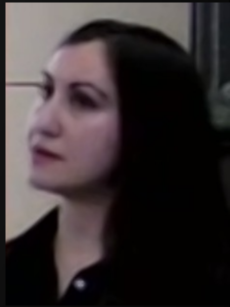
use the rubric for nonbinary evaluations

- Augment LLM with dumb ACC/PFC-like model
- Train dumb-PFC on past interactions, measure p(re-prompt), identify when it's time to switch from **fast to slow processing (thinking about thinking, system 2, cog control, etc)**
  - e.g., GPT4 nearperfect at identifying a response as toxic, but can't integrate this knowledge to not produce toxic content, dumb-PFC can reprompt & help
- Dumb-PFC can also decide when to
  - consult the internet or ground truth
  - recruit different skills/"personas"/attractor basins, e.g. to respond to the same question & take the best
- There can be different species of dumb-PFC (e.g., for different applications, Xbox vs. Bing vs. office/365 etc) Or multi-agent versions



Egner 2009, Nat Neuro

1. **DL needs PFC**: neuro, AI, behavior dims
2. **Transformer as HPC**: neuro, AI
3. **LLMs segment narrative structure**: AI, human-like behavior



"**executive functions** such as **planning** (Duncan, 1986), **abstract reasoning** (Donoso et al., 2014), **rule-learning** (Wallis et al., 2001), and **controlled** or **deliberate** processing (Miller & Cohen, 2001)"

PFC slows down for **top-down monitoring & control**: Memory & sequential planning (long-horizon), metacognition, orchestrating which regions should team up, increase communication, & or be more quiet ⇒ adapting the graph of functional connectivity to context & goals

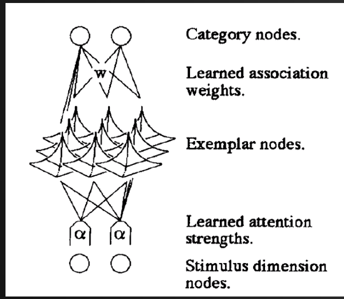
PFC

- to coordinate other processes & representations
- like in a multiagent constellation adaptive to task/ goals
- control as conductor of an orchestra

Nicholas Shea (london)

The importance of logical reasoning and its emergence in deep neural networks

1. representations in DNNs



Representing

- Implicitly, in a disposition to make transitions between representations:

  Space Needle → Seattle

- Explicitly:

  The Space Needle is in Seattle

BE REALIST ABOUT REPRESENTATIONS !

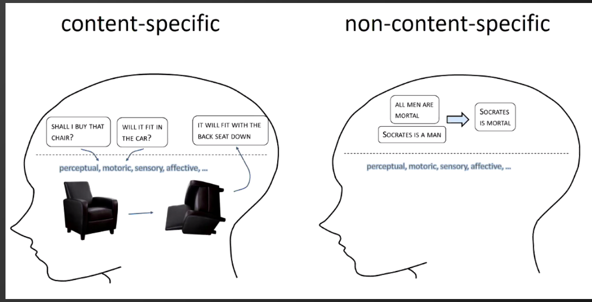
2. two types of representational transition (content-specific & non-content-specific)

Capacity for non-content-specific transitions is useful for:

(a) Inferences on representations far outside trained experience

(b) Inferences from stored explicit memories

3. humans: flexible reliance on both



4. hybrids in AI

Distinguish:

(a) Reasoning at output

(b) Internal non-content-specific computations

- (b) is unlikely:
  - (i) Patterns of errors, esp. out-of-distribution
  - (ii) What models do when trained specifically on logic: e.g. Traylor, Feiman & Pavlick (2021, ACL)

Potential hybrids

- LLM + reasoning engine ('tool use')
- LLM in two modes, via prompting

  E.g. 'Selection-inference':

  Cresswell, Shanahan & Higgins (2023, ICLR)

Non-content-specific transitions are useful for inferences on:

- Stored explicit memories
- Representations generated by general-purpose compositionality

Limitations:

- Computationally-demanding at decision time
- Frame problem / retrieval by relevance

are overcome by content-specific processing dispositions

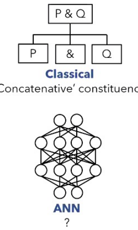
RAPHAËL MILLIÈRE (COLUMBIA)

COMPOSITIONALITY IN DEEP NEURAL NETWORKS

Principle of compositionality

"The meaning of a whole is a function of the meanings of the parts and of the way they are syntactically combined"
(Partee 1995)

A third way



DILEMMA

Human language and cognition are (largely) compositional

- If ANNs lack compositional representations with constituent structure, they *cannot* behave compositionally
- If ANNs have compositional representations with constituent structure, they merely *implement* a classical architecture

"Many current learning approaches are implicitly behaviorist in tint, ignoring the fact that the brain operates over representations that are organized into *structures* (not lists) based on compositional rules." (Marcus & Murphy 2022)

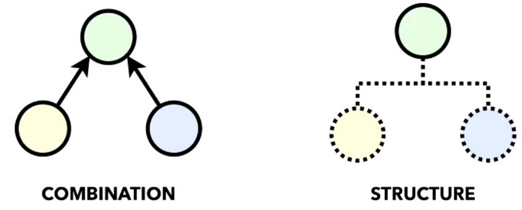
"It remains open that DNNs might mimic the performance of biological perception and cognition across a wide variety of domains and tasks by *implementing* core features of LoTs." (Quilty-Dunn et al. 2022)

"Do apparent successes of neural networks owe in part to implementing LoT-like structures, and if so, exactly what symbols and rules do they implement?" (Mandelbaum et al. 2022)

compositional behavior

INPUT	OUTPUT
A mat on a cat.	
Man bites dog. Who needs urgent care?	The dog

compositional representations



Compositional representations

"Compositionality is the classic idea that new representations can be constructed through the combination of primitive elements"
(Lake et al. 2016)

Conclusions

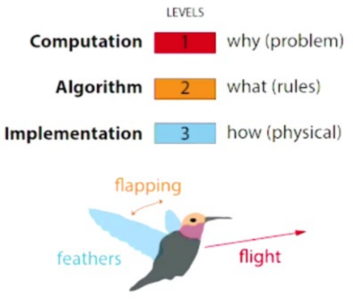
- DNNs can be given the resources to behave compositionally if they have the right features (biases, objective, size, data…)
- Functional compositionality in DNNs does not involve discrete constituent structure
- It provides a mechanism that approximates variable binding to varying degrees of precision
- Many open questions:
  - Architecture: is attention really special?
  - Augmentations: TPRs, parsers, explicit memory, logic engine…
  - Cognitive science: similar mechanisms in human cognition?

Grace Lindsay (NYU)

Developing neural systems understanding

1.) What kind of understanding do we seek? Does control demonstrate understanding?

- Many neuroscientists aim for something like an “algorithmic” level of understanding according to Marr’s framework

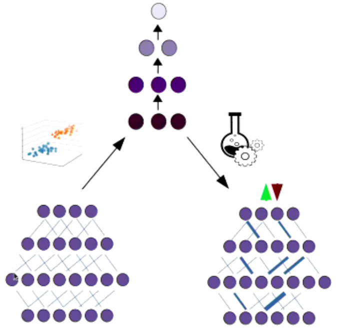


Marr (1982)

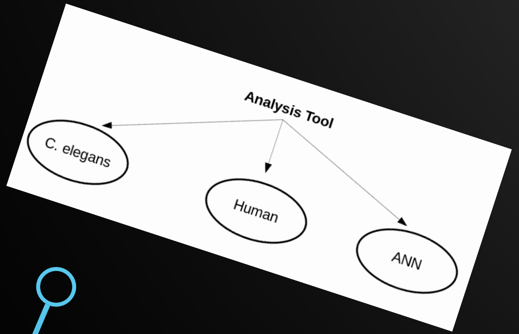
neurocritic

Plan: Operationalize it as “experimentally-validated understanding”

The application of an analysis method should result in a simplified understanding of the system. That understanding should be able to generate ideas for new experiments. Those experiments should verify the simplified model by demonstrating control over the behavior of the system



2.) What has to be true about two systems in order to be able to successfully apply a given analysis to both?



ANNs & brains share many features

- High dimensional
- Hierarchical/recurrent
- Nonlinear
- Distributed/Modular
- Task-optimized
- Information processing systems

3.) What will a successful ‘language’ for neural systems look like?

Development of ‘Neural Systems Understanding’

The development of this field **does not**:

- Require any specific claims about ANNs as models of the brain
- Assume that all neural systems should be submitted to the same tools
- Mean that all questions in neuroscience and AI can be solved with these methods

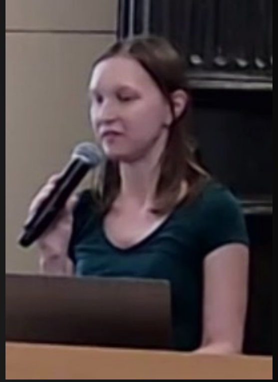
Symposium:
Linguistic and Cognitive Capacities of Large Language Models

Speakers

- Anna Ivanova (MIT)
- Nuhu Osman Attah (Pittsburgh)
- Patrick Butlin (Oxford)
- Philippe Verreault-Julien (Eindhoven)

Anna Ivanova (MIT)

Formal & functional competence in LLMs



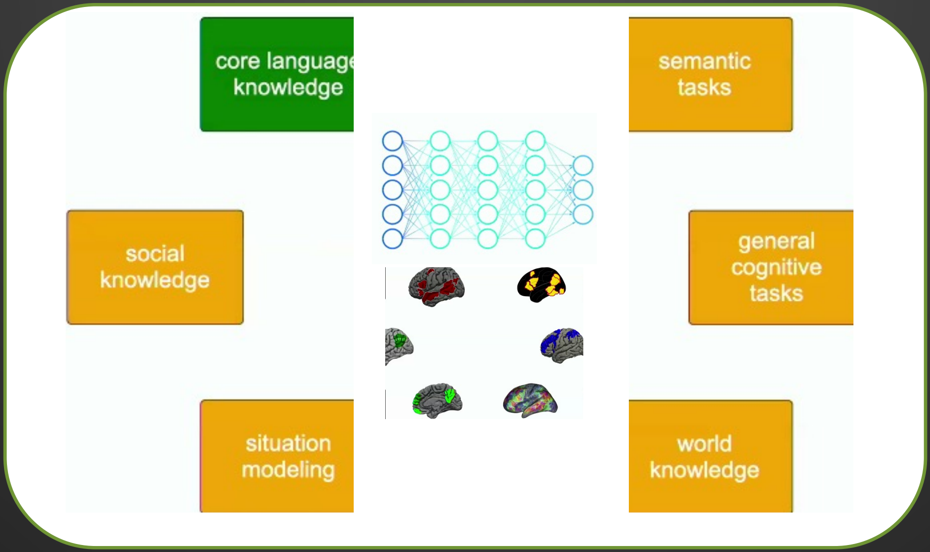
Fallacy #1

good at language → good at thought

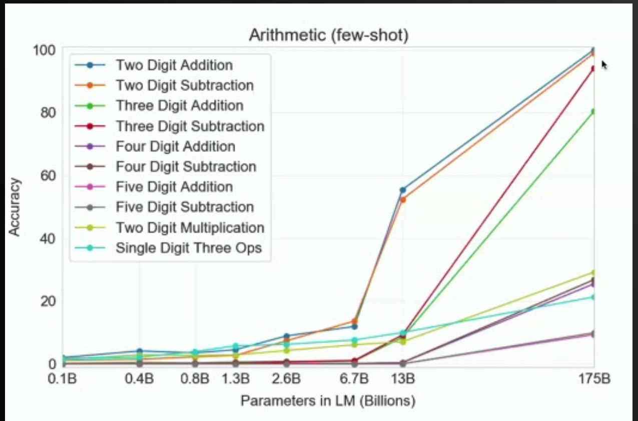
Fallacy #2

bad at language ← bad at thought

formal competences | functional competences



formal reasoning



Brown et al (2020)

Model	Overall
GPT-2	81.5
Human	88.6

davinci (175B)	0.84
GPT-NeoX (20B)	0.839
TNLG v2 (6.7B)	0.835
GPT-J (6B)	0.834

HELM

world knowledge

Language models learn a lot about the world. However, this knowledge is brittle, biased and incomplete.

The capital of Texas is **Austin**.
Boston? The capital of Texas is **Boston**.

Kassner & Schütze (2020)

- **Formal competence = knowledge of linguistic rules and patterns**
- **Functional competence = non-language-specific skills required for real-life language use**
- **This distinction (grounded in neuroscience) helps clarify the discourse around LLMs & suggests a way to build better language models.**

Nuhu Osman Attah (Pittsburgh)

Why think LLMs do not have any communicative intentions (CI) at all?

- Bender et al. 2021: because they don't have any mechanism to accommodate communicative intention nor are they trained to take such intentions into consideration

"It doesn't matter what internal mechanisms it uses, a sequence predictor is not, in itself, the kind of thing that could, even in principle, have communicative intent, and simply embedding it in a dialogue management system will not help." (Shanahan 2022).

But…

- Evidence suggests that attribution of intention (including self-attribution) is dependent on linguistic mastery – which suggests the semantics of intentional terms are significant for the ontogeny of communicative intention (Lohmann & Tomasello 2003).
- Moreover, the fine-tuning training phase of some Transformer LMs includes a dialogic component* (e.g. RLHF).

plausible mechanisms in LLMs

- In each case, the belief state representation is meant to estimate which of a set of possible effects a user intends to trigger.
- This representation is then used to guide a natural language generation module to take actions commensurate with this model of the user's intentions.
- This last point is important because everything I've said so far collapses the recognition/possession (of intention) distinction.

- Classical NLP systems would lend themselves positively to such a comparison.
- Until recent work, however, transformer-based LMs, might not have been thought to. It's an empirical matter whether they do.
- However, it is known that appropriate probes disentangle representational features in transformers which recapitulate the classic NLP pipeline, complete with distinct (hierarchical) representational sensitivity to parts of speech, semantic roles, and coreference (Tenney, Das, & Pavlick 2019, see also Clark et al. 2019).
- But if that is the case then the argumentative strategy of running through the system and trying to figure out intuitively where the representations of intentions might be encoded in it is dubious.

- If CI assumes Strong Griceanism, it won't get off the ground for all the well known reasons. (*So [out of our rhetorical magnanimity] let's assume it doesn't.*)
- Even if it attenuates its Gricean assumptions, it would still not be very convincing because…
  - Empirical parity.
  - There might be plausible mechanisms in LMs after all.
  - There might be more work for SL than CI Arguments suspect*.

Patrick Butlin (Oxford)

Can LLMs understand utterances?

previous claims:

1. Lack of *perception* of human environment does not prevent understanding
2. But lack of *functions or tasks* concerning this environment does

Butlin, P. (2021). Sharing Our Concepts with Machines.

NOW: function argument does not work

Claim: Understanding human utterances requires functions or tasks concerning the human environment

function argument

1. Understanding an utterance involves forming a representation with the same content
2. Content depends on function
   - A representation with the content *volcanoes erupt* has the function of carrying the information that volcanoes erupt
3. A system will only use information concerning the human environment if it has a function or task concerning that environment

Objection 1: Fine-tuning for new tasks

- Suppose a LM is fine-tuned to give correct answers to factual questions
- This is not a purely linguistic task
- It may use information about the human environment obtainable from its training data

Objection 2: Usefulness of information about the world

- Interpretability research sometimes posits representations with worldly content
- Hard to imagine how LMs produce some outputs without world knowledge

Objections 1 + 2: Discussion

System	Pretrained LM	Fine-tuned LM
Input	What is the capital of Estonia?	
Task	Provide a likely continuation of the text	Answer the question correctly
Information	'What is the capital of Estonia? Tallinn' is a relatively common string	Tallinn is the capital of Estonia

Two problems with this:

- The two facts are not independent, so features will carry both pieces of information
- Either piece of information could be used to perform either task

DETAILED ANALYSIS IS NEEDED TO CLARIFY REPRESENTATIONAL CONTENT IN LLMS

Philippe Verreault-Julien (Eindhoven)

Four Lessons LLMs teach us about understanding?

1. understanding comes in degrees
2. grasping matters
3. inferences aren’t the end of the story
4. understanding may not be compatible with lack of justification or falsehood

UNDERSTANDING

Threshold for understanding

Proto-understanding | Minimal understanding > Improved > Ideal understanding

INFERENCES

Abilities philosophers focus on are mostly **inferential**
LLMs are **good** (not perfect!) at inferences

- Counterfactual reasoning (e.g. Grimm 2006)
- Representation manipulation (Wilkenfeld 2013)
- Cognitive control (Hills 2016)

GRASPING MATTERS?

1. What are the constitutive abilities of grasping?
2. Is grasping phenomenal or inferential (Bourget 2017)?

Philosophers of understanding mostly:

a. Discuss whether some particular abilities are necessary for understanding
b. Endorse the inferential account

Grasping and its relationship to understanding may be crucial to establish whether LLMs understand

JUSTIFICATION OR FALSEHOOD

- **Is non-factive:** falsehood may afford understanding (Elgin 2017)
- **Doesn't require justification:** grasp of truth is sufficient (Dellsén 2017)