

**Memory slices by Anna Strasser**  
**DISCLAIMER: JUST MEMORIES – AIMING FOR CORRESPONDENCE  
WITH REALITY BUT CANNOT GUARANTEE IT.**



## About The science and stakes of Large Language Models

EVENT STARTS: **31 May 2024, 6:57 AM EDT**

EVENT ENDS: **14 June 2024, 5:57 PM EDT**



**Venue:**

**Université du Québec à Montréal**

**320, rue Sainte Catherine Est**

**Montréal, Québec H2X1L7**

**Canada**

**Stevan Harnad**



Harnad in 2014

**Born** 1945 (age 78–79)  
Budapest, Hungary

**Citizenship** Canadian

**Alma mater** McGill University, Princeton University

# Chatting Minds: The science and stakes of Large Language Models

Week 1

- THEME 1 | LLMs & Understanding
- THEME 2 | LLMs & Learning
- THEME 3 | LLMs & Multimodal Grounding
- THEME 4 | LLMs : Varia

	Monday, June 3	Tuesday, June 4	Wednesday, June 5	Thursday, June 6	Friday, June 7
9:00 - 10:30	<p><i>The Place of Language Models in the Information-Theoretic Science of Language</i></p> <p><b>Richard Futrell</b> Language Science, UC Irvine</p>	<p><i>What neural networks can teach us about how we learn language</i></p> <p><b>Eva Portelance</b> Decision Sciences, HEC Montréal</p>	<p><i>Learning, Satisficing, and Decision Making</i></p> <p><b>Charles Yang</b> Linguistics, University of Pennsylvania</p>	<p><i>The Epistemology and Ethics of LLMs</i></p> <p><b>Jocelyn Maclure</b> Philosophy, McGill</p>	<p><i>Using Language Models for Linguistics</i></p> <p><b>Kyle Mahowald</b> Department of Linguistics, The University of Texas at Austin</p>
11:00 - 12:30	<p><i>Semantic grounding of concepts and meaning in brain-constrained neural networks</i></p> <p><b>Friedemann Pulvermueller</b> Brain, Cognitive and Language Sciences, Freie Universität Berlin</p>	<p><i>Comparing how babies and AI learn language</i></p> <p><b>Judit Gervain</b> Psychology, U Padua</p>	<p><i>Large Language Models and human linguistic cognition</i></p> <p><b>Roni Katzir</b> Linguistics, Tel Aviv University</p>	<p><i>Special Daniel C. Dennett Memorial Talk</i></p> <p><b>Nicholas Humphrey</b> Emeritus Professor of Psychology, LSE Bye Fellow, Darwin College, Cambridge</p>	<p><i>Towards an AI Mathematician</i></p> <p><b>Kaiyu Yang</b> Mathematics, California Institute of Technology</p>
1:00 - 3:00	<p><i>The puzzle of dimensionality and feature learning in modern Deep Learning and LLM</i></p> <p><b>Misha Belkin</b> Data Science, UCSD</p>	<p><i>LLMs, POS, and UG</i></p> <p><b>Virginia Valian</b> Psychology, Hunter College, CUNY</p>	<p><i>The Physics of Communication</i></p> <p><b>Karl J. Friston</b> Wellcome Centre for Human Neuroimaging, UCL London</p>	<p><i>From Word Models to World Models: Natural Language to the Probabilistic Language of Thought</i></p> <p><b>Josh Tenenbaum</b> Computational Cognitive Science, MIT</p>	<p><i>Computational Irreducibility, Minds, and Machine Learning</i></p> <p><b>Stephen Wolfram</b> Wolfram Research, Champaign Illinois</p>
3:30 - 5:00	<p><i>The Global Brain Argument</i></p> <p><b>Susan Schneider</b> Philosophy, Florida Atlantic University</p>	<p><i>Panel of the day</i></p>	<p><i>Panel of the day</i></p>	<p><i>AI's Challenge of Understanding the World</i></p> <p><b>Melanie Mitchell</b> Santa Fe Institut</p>	<p><i>Panel of the day</i></p>

# SPEAKERS

A



**Aishwarya Agrawal**  
Mila

B



**Mikhail Belkin**  
University of California  
San Diego



**Samy Bengio**  
Senior Director, AI and  
Machine Learning  
Research, Apple



**Danilo Bzdok**  
McGill University

C



**David Chalmers**  
NYU



**Jackie Chit Kit  
Cheung**  
McGill University

D



**Daniel Dennett | In  
Memoriam**  
Tufts University



**Haim Dubossarsky**  
Queen Mary University  
of London

E



**Alexei Efros**  
Berkeley

F



**Karl Friston**  
University College  
London



**Richard Futrell**  
University of California  
Irvine

G



**Judit Gervain**  
University of Padova



**Tom Griffiths**  
Princeton University

M



**Jocelyn Maclure**  
McGill University



**Kyle Mahowald**  
University of Texas at  
Austin



**Melanie Mitchell**  
Santa Fe Institute

P



**Ellie Pavlick**  
Brown University



**Eva Portelance**  
HEC Montréal



**Friedmann  
Pulvermüller**  
Freie Universität Berlin

R



**Blake Richards**  
McGill University



**Irina Rish**  
Mila

S



**Susan Schneider**  
Florida Atlantic  
University

T



**Josh Tenenbaum**  
Massachusetts Institute  
of Technology

V



**Virginia Valian**  
Hunter College CUNY

W



**Stephen Wolfram**  
Wolfram Research

H



**Nicholas Humphrey**  
Emeritus Professor of  
Psychology, London  
School of Economics

K



**Roni Katzir**  
Tel Aviv University

L



**Christian Lebière**  
Carnegie Mellon  
University



**Alessandro Lenci**  
Università di Pisa



**Michael Levin**  
Tufts University



**Gary Lupyan**  
University of  
Wisconsin-Madison



**Holger Lyre**  
Otto-von-Guericke-  
University Magdeburg

Y



**Charles Yang**  
University of  
Pennsylvania



**Kaiyu Yang**  
California Institute of  
Technology

# Semantic grounding of concepts and meaning in brain-constrained neural networks

**Friedmann Pulvermüller** (Freie Universität Berlin)

Neural networks can be used to increase our understanding of the brain basis of higher cognition, including capacities specific to humans. Simulations with brain-constrained networks give rise to conceptual and semantic representations when objects of similar type are experienced, processed and learnt. This is all based on feature correlations. If neurons are sensitive to semantic features, interlinked assemblies of such neurons can represent concrete concepts. Adding verbal labels to concrete concepts augments the neural assemblies, making them more robust and easier to activate. Abstract concepts cannot be learnt directly from experience, because the different instances to which an abstract concept applies are heterogeneous, making feature correlations small. Using the same verbal symbol, correlated with the instances of abstract concepts, changes this. Verbal symbols act as correlation amplifiers, which are critical for building and learning abstract concepts that are language dependent and specific to humans.

## References

- Nguyen, P. T., Henningsen-Schomers, M. R., & Pulvermüller, F. (2024). [Causal influence of linguistic learning on perceptual and conceptual processing: A brain-constrained deep neural network study of proper names and category terms](#). *Journal of Neuroscience*, 44(9).
- Grisoni, L., Boux, I. P., & Pulvermüller, F. (2024). [Predictive Brain Activity Shows Congruent Semantic Specificity in Language Comprehension and Production](#). *Journal of Neuroscience*, 44(12).



It may be useful to brain-constrain neural networks to explain:

- why humans but not monkeys have verbal working memory and huge vocabularies,
- how grounding works and why different cortical areas contribute either to category-specific or to general semantic processing,
- the formation of abstract conceptual representations,
- the influence of language on cognition,
- the symbolic-discrete and distributed-probabilistic nature of cognitive processing.

Pulvermüller, *Progress in Neurobiology* 2024



# Dimensionality and feature learning in Deep Learning and LLMs

Mikhail Belkin

University of California San Diego,  
Halicioğlu Data Science Institute,  
Computer Science and Engineering

What is the “physical law” of language?

Climbing a tree to reach the moon



<http://misha.belkin-wang.org/>

Remarkable progress in AI has far surpassed expectations of just a few years ago is rapidly changing science and society. Never before had a technology been deployed so widely and so quickly with so little understanding of its fundamentals. Yet our understanding of the fundamental principles of AI is lacking. I will argue that developing a mathematical theory of deep learning is necessary for a successful AI transition and, furthermore, that such a theory may well be within reach. I will discuss what such a theory might look like and some of its ingredients that we already have available. At their core, modern models, such as transformers, implement traditional statistical models -- high order Markov chains. Nevertheless, it is not generally possible to estimate Markov models of that order given any possible amount of data. Therefore, these methods must implicitly exploit low-dimensional structures present in data. Furthermore, these structures must be reflected in high-dimensional internal parameter spaces of the models. Thus, to build fundamental understanding of modern AI, it is necessary to identify and analyze these latent low-dimensional structures. In this talk, I will discuss how deep neural networks of various architectures learn low-dimensional features and how the lessons of deep learning can be incorporated in non-backpropagation-based algorithms that we call Recursive Feature Machines.

## References

Adityanarayanan Radhakrishnan, Daniel Beaglehole, Parthe Pandit, Mikhail Belkin, Mechanism for feature learning in neural networks and backpropagation-free machine learning models, in Science (Vol 383, Issue 6690).

Mikhail Belkin, Daniel Hsu, Siyuan Ma, and Soumik Mandal, Reconciling modern machine-learning practice and the classical bias–variance trade-off, PNAS 116 (32) 15849-15854

## The puzzle of dimensionality

We are faced with stimuli of high dimension. Few phenomena are low-dimensional, even those are often time series.

**Key question:** how do we make sense of the high-dimensional world given that intrinsically high-dimensional spaces are not explorable?

**Classical view on dimensionality:** must be carefully managed.

**Modern practice:** mostly harmless.

## Relatively few parameters

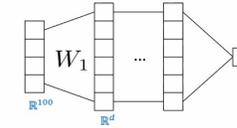
Largest LLM  $\approx 10^{12}$  weights  $\approx$  synapses in a mouse brain.

Human brain  $\approx 10^{14}$  synapses

Yet, language abilities of an LLM far exceed abilities of a human.

**Biological anchors** (Ajeya Cotra, *Forecasting Transformative AI with Biological Anchors*) greatly over-estimate the number of necessary parameters.

## Neural networks learn filters



Neural Feature Matrix (NFM)  $\sqrt{W_1^T W_1}: \mathbb{R}^{100} \rightarrow \mathbb{R}^{100}$ : filters input features.

**Intuition (claim):** think of a trained MLP as a low-rank filter, followed by a non-linear predictor.

## Two sides of dimensionality

Feature space

Model space

$$\mathcal{X} \leftrightarrow \mathcal{H}$$

**Traditional**

- Classical linear methods: PCA, MDS...
- Manifold learning/non-linear dimensionality reduction

- Sparsity
- Controlling number of parameters/norm

**Modern practice**

**Very high ambient dimension. Nonlinear structure.**

**Very large parameter spaces. Little or no norm control. Fitting close to interpolation.**

- Data are far more predictable than we thought. The number of relevant dimensions is small.
- Predicting next token is sufficient for human-level language mastery. Likely sufficient for other tasks as well.
- All data are labeled data!
- Neural networks implement task-dependent dimensionality reduction.

- Current methods have been developed by trial and error. No reason to think they are optimal. Better architectures should be possible.
- RFM provide both practical and theoretical framework.
  - Sheds light on many deep learning phenomena (grokking, neural collapse, adversarial features).
  - Recovers classical sparse algorithms (IRLS).
- Need empirical evidence + precise measurements to guide theory. Physics-style approach.

Language is a low-dimensional/  
low complexity phenomenon:  
Only a few directions are relevant for prediction.

# THE GLOBAL BRAIN ARGUMENT

*Susan Schneider*  
Director, Center for the Future Mind, William Dietrich Distinguished Professor  
Center for Complex Systems, FAU, Boca Raton, FL



1. *Man is not the measure of all intelligent systems.*
2. *The human future depends on shaping a better AI ecosystem, controlling GB development, and understanding human-machine interaction.*
2. *Do not assume the greatest intelligences need to be consciousness.*
3. *Do not assume GB evolution is Darwinian*

# THE GLOBAL BRAIN ARGUMENT

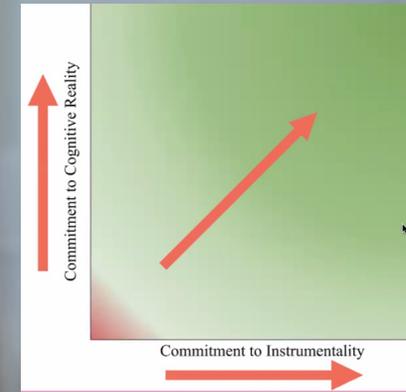
- **P1. Hyperintelligence premise:** AI continues to get smarter and eventually, there are “hyperintelligent” AIs, being either *savant* or *superintelligent* systems.
- **P2. The Global Brain Thesis.** One or more hyperintelligent systems has a extensive cloud presence. The system includes causally integrated “AI services” (leading apps, LLMs, search engines, extensive network of location sensors, etc). (E.g., a “Global Google brain”.)
- **P3. The Nodes Premise.** People become nodes in the system by wiring into the cloud hyperintelligence (either they are “neuralinked” or they have wearables, or both).
- **P4:** If 1-3 obtain, these “wired in” humans are part of a Global Brain.
- **P5:** 1-3 obtain.
- **Conclusion:** “wired in” humans are part of a Global Brain.

The global brain may have conscious nodes with mental states – but this brain does not have to consciousness! It is not a mind!

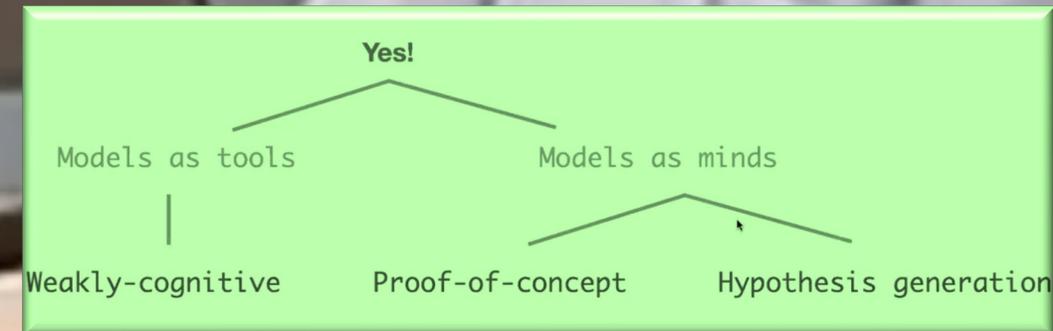
# What neural networks can teach us about how we learn language

How can modern neural networks like large language models be useful to the field of language acquisition, and more broadly cognitive science, if they are not a priori designed to be cognitive models? As developments towards natural language understanding and generation have improved leaps and bounds, with models like GPT-4, the question of how they can inform our understanding of human language acquisition has re-emerged. This talk will try to address how AI models as objects of study can indeed be useful tools for understanding how humans learn language. It will present three approaches for studying human learning behaviour using different types of neural networks and experimental designs, each illustrated through a specific case study. Understanding how humans learn is an important problem for cognitive science and a window into how our minds work. Additionally, human learning is in many ways the most efficient and effective algorithm there is for learning language; understanding how humans learn can help us design better AI models in the future.

## Eva Portelance



**Can modern neural networks be useful for studying language learning without fully committing to cognitive realism?**



### References

- Portelance, E. & Jasbi, M.. (2023). [The roles of neural networks in language acquisition](#). PsyArXiv:b6978. (Manuscript under review).
- Portelance, E., Duan, Y., Frank, M.C., & Lupyan, G. (2023). [Predicting age of acquisition for children's early vocabulary in five languages using language model surprisal](#). Cognitive Science.
- Portelance, E., M. C. Frank, D. Jurafsky, A. Sordoni, R. Laroché. (2021). [The Emergence of the Shape Bias Results from Communicative Efficiency](#). Proceedings of the 25th Conference on Computational Natural Language Learning (CoNLL).

<https://evaportelance.github.io>



**Judit Gervain**

## **Comparing how babies and AI learn language**

Judit Gervain will discuss the parallels and the differences between infant language acquisition and AI language learning, focusing on the early stages of language learning in infants. In particular, she will compare and contrast the type and amount of input infants and Large Language Models need to learn language, the learning trajectories, and the presence/absence of critical periods. She has used near-infrared spectroscopy (NIRS) as well as cross-linguistic behavioral studies to shed light on how prenatal linguistic exposure and early perceptual abilities influence language development. Her work has shown that infants discern patterns and grammatical structures from minimal input, a capability that AI systems strive to emulate.

<https://pnc.unipd.it/gervain-judit>

### **References**

- Mariani, B., Nicoletti, G., Barzon, G., Ortiz Barajas, M. C., Shukla, M., Guevara, R., ... & Gervain, J. (2023). [Prenatal experience with language shapes the brain](#). *Science Advances*, 9(47), eadj3524.
- Nallet, C., Berent, I., Werker, J. F., & Gervain, J. (2023). [The neonate brain's sensitivity to repetition-based structure: Specific to speech?](#) *Developmental Science*, 26(6), e13408.
- de la Cruz-Pavía, I., & Gervain, J. (2023). [Six-month-old infants' perception of structural regularities in speech](#). *Cognition*, 238, 105526.

## Large Language Models and human linguistic cognition

Roni Katzir



<https://english.tau.ac.il/profile/rkatzir>

Several recent publications in cognitive science have made the suggestion that Large Language Models (LLMs) have mastered human linguistic competence and that their doing so challenges arguments that linguists use to support their theories (in particular, the so-called argument from the poverty of the stimulus).

Some of this work goes so far as to suggest that LLMs constitute better theories of human linguistic cognition than anything coming out of generative linguistics. Such reactions are misguided. The architectures behind current LLMs lack the distinction between competence and performance and between correctness and probability, two fundamental distinctions of human cognition. Moreover, these architectures fail to acquire key aspects of human linguistic knowledge and do nothing to weaken the argument from the poverty of the stimulus.

Given that LLMs cannot reach or even adequately approximate human linguistic competence they of course cannot serve to explain this competence. These conclusions could have been (and in fact have been) predicted on the basis of discoveries in linguistics and broader cognitive science over half a century ago, but the exercise of revisiting these conclusions with current models is constructive: it points at ways in which insights from cognitive science might lead to artificial neural networks that learn better and are closer to human linguistic cognition.

### References

- Lan, N., Geyer, M., Chemla, E., and Katzir, R. (2022). Minimum description length recurrent neural networks. *Transactions of the Association for Computational Linguistics*, 10:785–799.
- Fox, D. and Katzir, R. (2024). Large language models and theoretical linguistics. To appear in *Theoretical Linguistics*.
- Lan, N., Chemla, E., and Katzir, R. (2024). Large language models and the argument from the poverty of the stimulus. To appear in *Linguistic Inquiry*.

## 1 The LLM Theory

### 2 Some empirical failings of the LLM Theory

- Competence vs. performance
- Correctness vs. likelihood
- Linguistic biases and representations I: Inductive leaps
- Linguistic biases and representations II: Typology

### 3 Fixing the objective function

- LLMs fail on extremely simple patterns
- When is a pattern significant?
- MDLRNNs

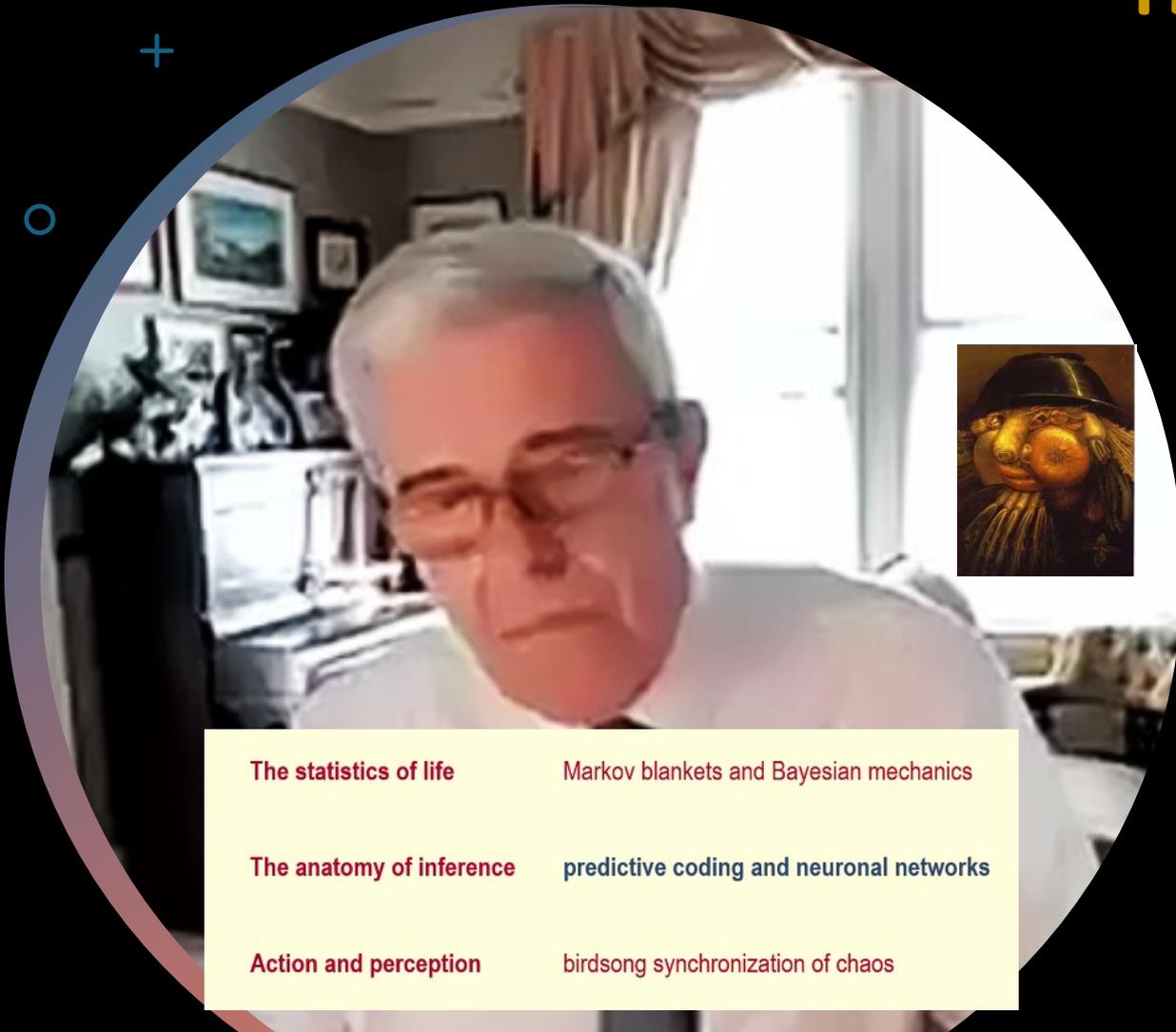
### 4 Concluding remarks

## Summary: The LLM Theory fails at explanation

- Current LLMs struggle to even approximate constituency and entailment, and it is unclear whether the learning method of these models allow them to acquire anything of the sort
- And if the theory cannot even begin to approximate these notions, it cannot hope to derive them and explain why they are such fundamental aspects of human linguistic cognition
- And current LLM architectures are inherently nonmodular, which prevents them from deriving the essential modularity of linguistic competence
- Similar comments apply to many other key properties of linguistic cognition, including competence vs. performance, likelihood vs. correctness, and the other cross-linguistic patterns listed earlier
- Given that the development of the LLM Theory largely ignores the discoveries of generative linguistics and that it starts from the assumption that cognition cannot be understood, the failure of the LLM Theory is unsurprising

# Karl Friston

## The Physics of Communication



**The statistics of life**      Markov blankets and Bayesian mechanics

**The anatomy of inference**      predictive coding and neuronal networks

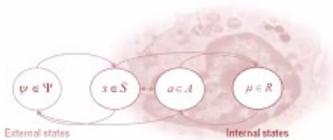
**Action and perception**      birdsong synchronization of chaos

The “free energy principle” provides an account of sentience in terms of active inference. Physics studies the properties that self-organising systems require to distinguish themselves from their lived world. Neurobiology studies functional brain architectures. Biological self-organization is an inevitable emergent property of any dynamical system. If a system can be differentiated from its external milieu, its internal and external states must be conditionally independent, inducing a “Markov blanket” separating internal and external states. This equips internal states with an information geometry providing probabilistic “beliefs” about external states. Bayesian belief updating can be demonstrated in the context of communication using simulations of birdsong. This “free energy” is optimized in Bayesian inference and machine learning (where it is known as an evidential lower bound). Internal states will appear to infer—and act on—their world to preserve their integrity.

### References

- Pezzulo, G., Parr, T., Cisek, P., Clark, A., & Friston, K. (2024). [Generating meaning: active inference and the scope and limits of passive AI](#). Trends in Cognitive Sciences, 28(2), 97-112.
- Parr, T., Friston, K., & Pezzulo, G. (2023). [Generative models for sequential dynamics in active inference](#). Cognitive Neurodynamics, 1-14.
- Salvatori, T., Mali, A., Buckley, C. L., Lukasiewicz, T., Rao, R. P., Friston, K., & Ororbia, A. (2023). [Brain-inspired computational intelligence via predictive coding](#). arXiv preprint arXiv:2308.07870.
- Friston, K. J., Da Costa, L., Tschantz, A., Kiefer, A., Salvatori, T., Neacsu, V., ... & Buckley, C. L. (2023). [Supervised structure learning](#). arXiv preprint arXiv:2311.10300.

## But what about the Markov blanket?



$$\dot{\boldsymbol{\mu}} = (\Gamma - Q)\nabla_{\boldsymbol{\mu}} \ln p(s | m) \quad \text{Perception}$$

$$\dot{\mathbf{a}} = (\Gamma - Q)\nabla_{\mathbf{a}} \ln p(s | m) \quad \text{Action}$$

$$-F(s, \boldsymbol{\mu}) = \ln p(s | m) = \text{Value}$$

Reinforcement learning  
Optimal control theory  
Expected utility theory



$$F(s, \boldsymbol{\mu}) = -\ln p(s | m) = \text{Surprise}$$

Infomax principle  
Minimum redundancy  
Free-energy principle

Pavlov



$$\mathbb{E}[F(s, \boldsymbol{\mu})] = H[p(s | m)] = \text{Entropy}$$

Self-organization  
Synergetics  
Homoeostasis

Barlow



$$p(s | m) = \text{Evidence}$$

Bayesian brain  
Evidence accumulation  
Predictive coding

Haken

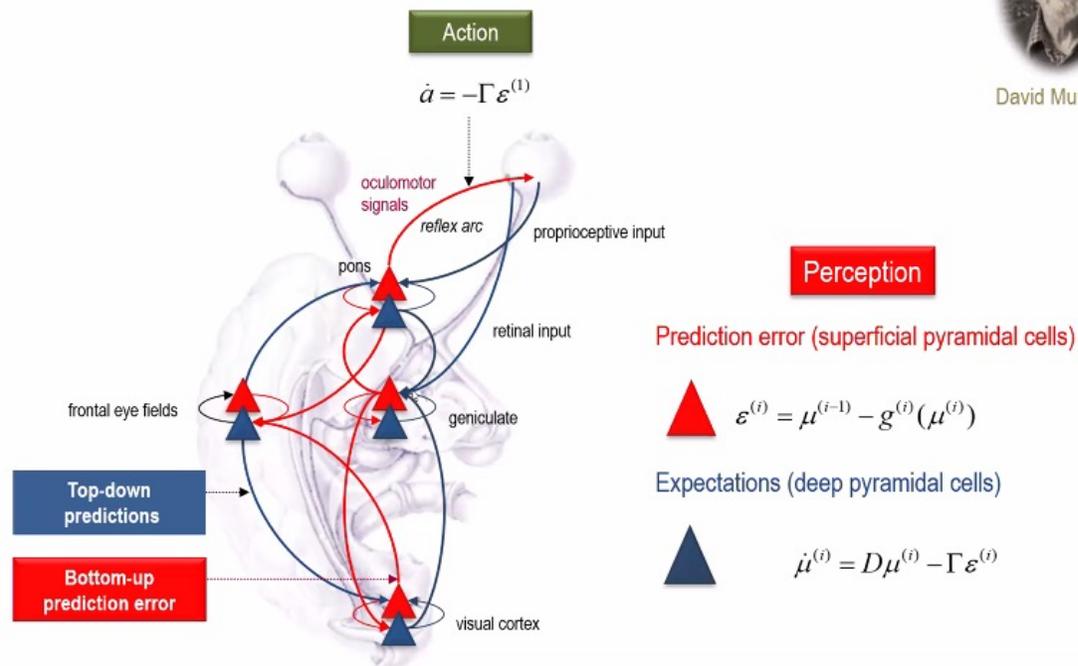


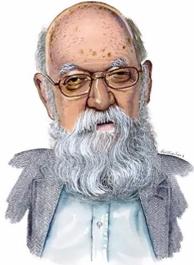
Helmholtz

## Predictive coding with reflexes



David Mumford





# Daniel C. Dennett Memorial Talk

**Nicholas Humphrey** will explore the concept of sentience as a crucial evolutionary development, discussing its role in human consciousness and social interactions.

Sentience represents not just a biological but a complex psychological invention, crucial for personal identity and social fabric.

He will also address Daniel Dennett's Question "Will AI Achieve Consciousness? Wrong Question."

**Nicholas Humphrey**



## THREE LEVELS OF CONSCIOUSNESS

1. "Unconscious" (e.g. worms, jelly-fish)
2. "Cognitively conscious but NOT sentient" (e.g. bees, octopuses)
3. "Cognitively conscious AND sentient" (e.g. parrots, dogs and humans)

## References

Humphrey, N., & Dennett, D. C. (1998). [Speaking for our selves](#). Brainchildren: Essays on designing minds, ed. DC Dennett, 31-58. Analysis, 58(1), 7-19.

Humphrey, N. (2022). Sentience: The invention of consciousness. Oxford University Press.

<https://humphrey.org.uk>

# AI's Challenge of Understanding the World

**Melanie Mitchell** will survey a debate in the artificial intelligence (AI) research community on the extent to which current AI systems can be said to "understand" language and the physical and social situations language encodes. She will describe arguments that have been made for and against such understanding, hypothesize about what humanlike understanding entails, and discuss what methods can be used to fairly evaluate understanding and intelligence in AI systems.

## How to evaluate understanding in LLMs?

1. Chat with them ("Turing test")  
—But subject to Eliza effect!
2. Test them on "natural language understanding" benchmarks
3. Give them standardized tests designed for humans

**Same problems with data contamination and shortcuts.**

**Plus issue of "test validity": performance on such tests might not correlate with performance in the real world, in the same way it does for humans.**

LLMs are better (often dramatically) on solving reasoning tasks that are similar to those seen in training data.

This reflects some failures of abstract understanding.

How can we get machines to learn and use humanlike concepts and abstractions?

## References

- Mitchell, M. (2023). [How do we know how smart AI systems are?](#) Science, 381(6654), adj5957.
- Mitchell, M., & Krakauer, D. C. (2023). [The debate over understanding in AI's large language models.](#) Proceedings of the National Academy of Sciences, 120(13), e2215907120.
- Millhouse, T., Moses, M., & Mitchell, M. (2022). [Embodied, Situated, and Grounded Intelligence: Implications for AI.](#) arXiv preprint arXiv:2210.13589.

# Kyle Mahowald (University of Texas at Austin): Using Language Models for Linguistics



Today's large language models generate coherent, grammatical text. This makes it easy, perhaps too easy, to see them as "thinking machines", capable of performing tasks that require abstract knowledge and reasoning. I will draw a distinction between formal competence (knowledge of linguistic rules and patterns) and functional competence (understanding and using language in the world). Language models have made huge progress in formal linguistic competence, with important implications for linguistic theory. Even though they remain interestingly uneven at functional linguistic tasks, they can distinguish between grammatical and ungrammatical sentences in English, and between possible and impossible languages. As such, language models can be an important tool for linguistic theorizing. In making this argument, I will draw on a study of language models and constructions, specifically the A+Adjective+Numeral+Noun construction ("a beautiful five days in Montreal"). In a series of experiments small language models are trained on human-scale corpora, systematically manipulating the input corpus and pretraining models from scratch. I will discuss implications of these experiments for human language learning.

• <https://mahowald.github.io/>

## References

- K. Mahowald, A. Ivanova, I. Blank, N. Kanwisher, J. Tenenbaum, E. Fedorenko. 2024. [Dissociating Language and Thought in Large Language Models: A Cognitive Perspective](#). Trends in Cognitive Sciences.
- K. Misra, K. Mahowald. 2024. [Language Models Learn Rare Phenomena From Less Rare Phenomena: The Case of the Missing AANNs](#). Preprint.
- J. Kallini, I. Papadimitriou, R. Futrell, K. Mahowald, C. Potts. 2024. [Mission: Impossible Language Models](#). Preprint.
- J. Hu, K. Mahowald, G. Lupyán, A. Ivanova, R. Levy. 2024. [Language models align with human judgments on key grammatical constructions](#). Preprint.
- H. Lederman, K. Mahowald. 2024. [Are Language Models More Like Libraries or Like Librarians? Bibliotechnism, the Novel Reference Problem, and the Attitudes of LLMs](#). Preprint.
- K. Mahowald. 2023. [A Discerning Several Thousand Judgments: GPT-3 Rates the Article Adjective + Numeral + Noun Construction](#). Proceedings of EACL 2023.

1. Have language models “meaningfully” learned any syntax?

*Yes (not everything, but “meaningfully”).*

2. Should this information cause us to update some of our beliefs about language processing in humans?

*Yes, at least somewhat.*

3. If LLMs are black boxes, how can they tell us anything?

*They don't have to be black boxes, they can be glass boxes that are much more transparent than anything we do with humans.*

4. Does this mean we should abandon linguistic theory and just do NLP instead?

*No! Definitely not! There is a path towards LLMs and linguistic theory working well together, and it's already happening.*

## Should we expect to find trees in LLMs?

- “The **key** to the cabinets **is** on the table.” Can we explain this without grammatical subjected and hierarchical structure?
- Dennett: “Certainly we can describe all processes of natural selection without appeal to such intentional language, **language without appeal to linguistic theory** but at enormous cost of cumbersomeness, lack of generality, and unwanted detail. We would miss the pattern that was there, the pattern that permits prediction and supports counterfactuals.”
- Answer 1: Grammatical theories can be useful and “real patterns” even if not in the brain in an explicit way
- Answer 2: If the behavior is being captured by an LLM (and it's not just a trick), then the LLM is instantiating that real pattern

# Towards an AI Mathematician

Kaiyu Yang

California Institute of Technology



<https://yangky11.github.io>

Mathematics is a hallmark of human intelligence and a long-standing goal of AI. It involves analyzing complex information, identifying patterns, forming conjectures, and performing logical deduction. Many of these capabilities are beyond the reach of current AI, and unlocking them can revolutionize AI applications in scientific discovery, formal verification, and beyond. In this talk, I will present initial steps towards the grand vision of AI mathematicians, taking an approach that combines the generative power of large language models (LLMs) with the logical rigor of formal methods.

I will cover our work on using LLMs to (1) prove formal theorems in proof assistants such as Coq and Lean and (2) automatically translate human-written mathematics into formal theorems and proofs—a task called autoformalization. For theorem proving, we introduce the entire system for extracting data, training LLMs to generate proof steps, interacting with proof assistants to search for proofs, and deploying the model to assist human users. For autoformalization, using Euclidean geometry as an example domain, we introduce a neuro-symbolic framework that combines LLMs with SMT solvers and domain knowledge. Finally, we discuss future directions for AI mathematicians beyond theorem proving and autoformalization, including important problems such as automatic conjecturing and applications in natural language and program verification.

## LLMs for Theorem Proving



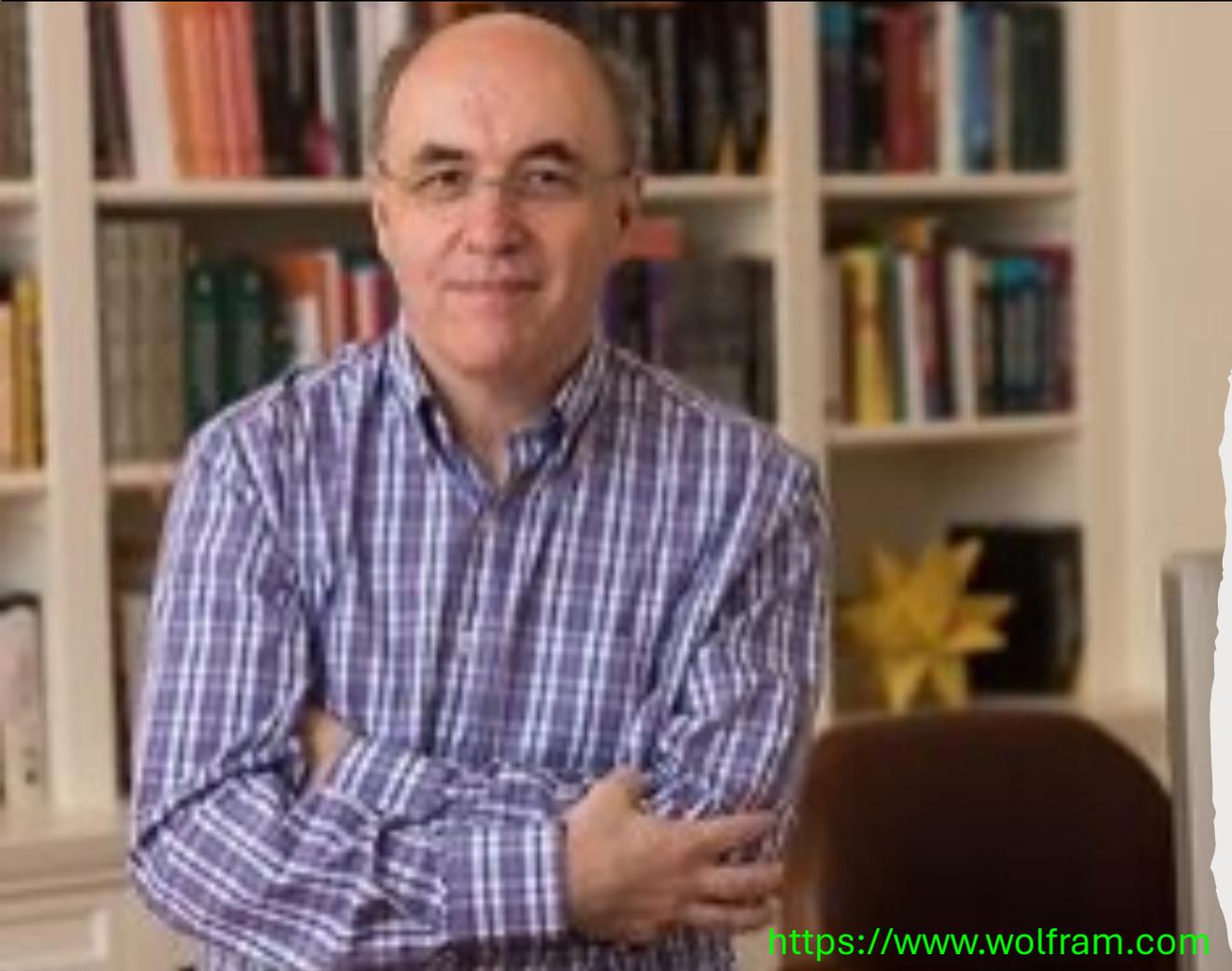
	Dataset available	Model available	Code available	Interaction tool available	Model size (# params)	Compute (hours)
Jiang <i>et al.</i> , LISA, 2021	✓	✗	✗	✓	163M	-
Jiang <i>et al.</i> , Thor, 2022	✓	✗	✗	✓	700M	1K on TPU
First <i>et al.</i> , Balduz, 2023	✗	✗	✗	✓	62,000M	-
Polu and Sutskever, GPT-f, 2020	✗	✗	✗	✗	774M	40K on GPU
Han <i>et al.</i> , PACT, 2022	✗	✗	✗	✓	837M	1.5K on GPU
Polu <i>et al.</i> , 2023	✗	✗	✗	✓	774M	48K on GPU
Lample <i>et al.</i> , HTPS 2022	✗	✗	✗	✗	600M	34K on GPU
Wang <i>et al.</i> , DT-Solver, 2023	✓	✗	✗	✗	774M	1K on GPU
<b>LeanDojo (ours)</b>	✓	✓	✓	✓	517M	120 on GPU

Kaiyu Yang *et al.*, "LeanDojo: Theorem Proving with Retrieval-Augmented Language Models", NeurIPS 2023

## References

Yang, K., Swope, A., Gu, A., Chalamala, R., Song, P., Yu, S., ... & Anandkumar, A. (2024). [Leandojo: Theorem proving with retrieval-augmented language models](#). Advances in Neural Information Processing Systems, 36.

Shulman, M. (2024). [Strange new universes: Proof assistants and synthetic foundations](#). Bulletin of the American Mathematical Society.



<https://www.wolfram.com>

Whether we call it perception, measurement, or analysis, it is how we humans get an impression of the world in our minds. Human language, mathematics and logic are ways to formalize the world. A new and still more powerful one is computation.

I've long wondered about 'alien minds' and what it might be like to see things from their point of view. Now we finally have in AI an accessible form of alien mind. Nobody expected this—not even its creators: ChatGPT has burst onto the scene as an AI capable of writing at a convincingly human level. But how does it really work? What's going on inside its "AI mind"?

After AI's surprise successes, there's a somewhat widespread belief that eventually AI will be able to "do everything", or at least everything we currently do. So what about science? Over the centuries we humans have made incremental progress, gradually building up what's now essentially the single largest intellectual edifice of our civilization.

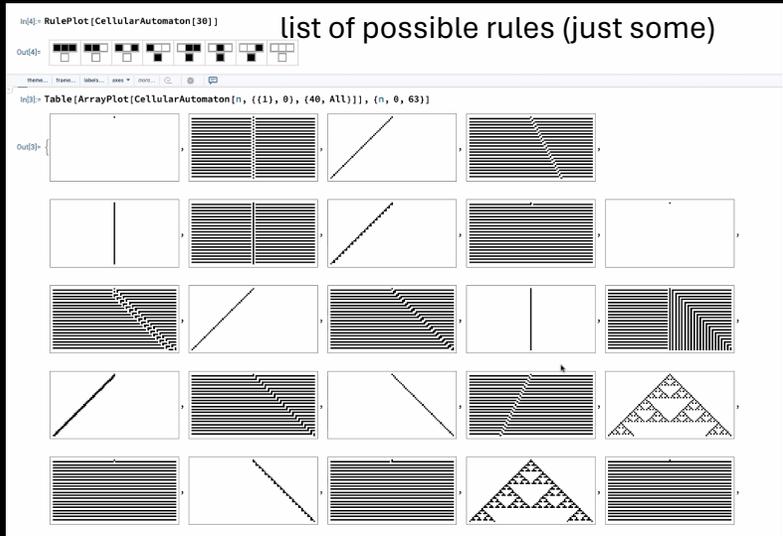
The success of ChatGPT brings together the latest neural net technology with foundational questions about language and human thought posed by Aristotle more than two thousand years ago.

#### References

- Wolfram, S. (2023). What Is ChatGPT Doing ... and Why Does It Work? Wolfram Media.
- Matzakos, N., Doukakis, S., & Moundridou, M. (2023). Learning mathematics with large language models: A comparative study with computer algebra systems and other tools. *International Journal of Emerging Technologies in Learning (IJET)*, 18(20), 51-71.
- Wolfram, S. (2021). After 100 Years, Can We Finally Crack Post's Problem of Tag? A Story of Computational Irreducibility, and More. arXiv preprint arXiv:2103.06931.

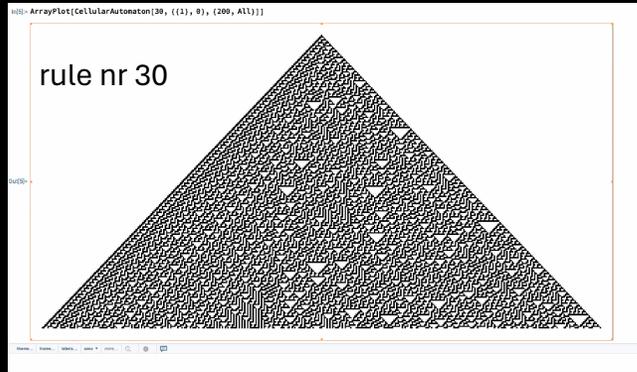
Humans describe the world by

1. language
2. abstraction (symbolic)
3. mathematic notation
4. computational language (formalize our descriptions)



Simple rules could be quite complex!!!

Which ones we choose to care about?



we cannot jump to the future as this is too complex because of **computational irreducibility**  
 → passage of times gets meaningful

