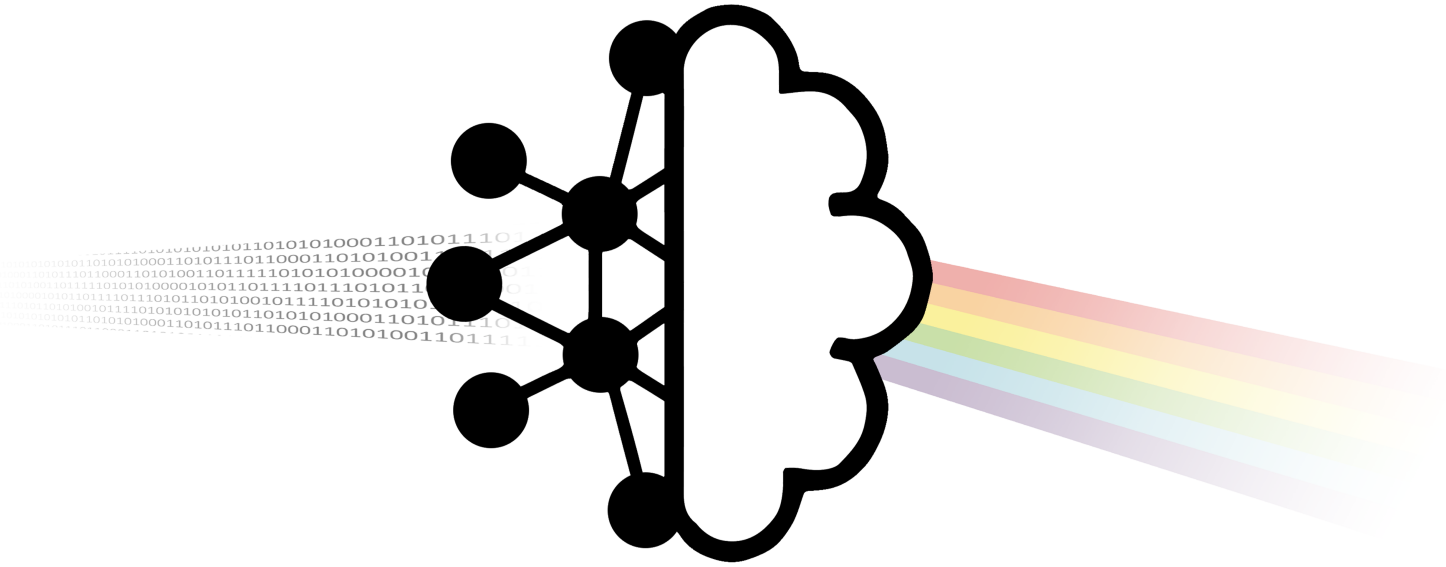
# The Philosophy of Deep Learning

**March 24-26 2023 · New York University**



**phildeeplearning.github.io**

COLUMBIA UNIVERSITY
Presidential Scholars in Society and Neuroscience

NYU CENTER FOR MIND, BRAIN, AND CONSCIOUSNESS

This is the short version of the booklet for print use.
For more information about the conference, registration and livestream, see:
https://phildeeplearning.github.io/

# Contents

**About** **4**
Registration . . . . . 4
Livestream . . . . . 4

**Timetable** **5**
Friday, March 24th – Pre-Conference Debate . . . . . 5
Saturday, March 25th – Conference Day 1 . . . . . 6
Sunday, March 26th – Conference Day 2 . . . . . 8

**List of Abstracts – Talks** **9**
Saturday, March 25th . . . . . 9
Sunday, March 26th – Conference Day 2 . . . . . 12

**List of Posters** **17**

**Useful Information** **23**

**Sponsors** **24**

# About

The Philosophy of Deep Learning is a two-day conference (March 25-26th) plus pre-conference debate (March 24th) on the philosophy of deep learning, organized by Ned Block (New York University), David Chalmers (New York University) and Raphaël Millière (Columbia University).

The conference will explore current issues in AI research from a philosophical perspective, with particular attention to recent work on deep artificial neural networks. The goal is to bring together philosophers and scientists who are thinking about these systems in order to gain a better understanding of their capacities, their limitations, and their relationship to human cognition.

The conference will focus especially on topics in the philosophy of cognitive science (rather than on topics in AI ethics and safety). It will explore questions such as:

- What cognitive capacities, if any, do current deep learning systems possess?
- What cognitive capacities might future deep learning systems possess?
- What kind of representations can we ascribe to artificial neural networks?
- Could a large language model genuinely understand language?
- What do deep learning systems tell us about human cognition, and vice versa?
- How can we develop a theoretical understanding of deep learning systems?
- How do deep learning systems bear on philosophical debates such as rationalism vs empiricism and classical vs. nonclassical views of cognition.
- What are the key obstacles on the path from current deep learning systems to human-level cognition?

A pre-conference debate on Friday, March 24th will tackle the question “Do large language models need sensory grounding for meaning and understanding?”.

## Registration

Attendance is free but requires registration. Please register in advance at phildeeplearning.github.io. Note that an Eventbrite ticket does not guarantee a seat and that the venue may be at capacity if you do not arrive early.

## Livestream

The conference will be livestreamed and recorded via Zoom on phildeeplearning.github.io/streaming. Please note that this is not a hybrid conference, and viewers on the livestream will not be able to ask questions or participate remotely.

# Timetable

## Friday, March 24th – Pre-Conference Debate

**Cantor Film Center, Room 200, 36 East 8th Street**

| | | | |
|---|---|---|---|
| 5:30–7:30 | **Do Language Models Need Sensory Grounding for Meaning and Understanding?** | | |
| **YES** | **Yann LeCun**<br>NYU/Meta AI | **Brenden Lake**<br>NYU | **Jacob Browning**<br>NYU |
| **NO** | **Ellie Pavlick**<br>Brown/Google AI | **David Chalmers**<br>NYU | **Gary Lupyan**<br>Wisconsin |
| 7:30–8:30 | **Reception (Silverstein Lounge, 32 Waverly Place)** | | |

# Saturday, March 25th – Conference Day 1

**19 West 4th Street, Room 101**

ML: Main Lecture; ST: Symposium Talk; PC: Panel Contribution; PP: Poster Presentation.

| | | | |
|---|---|---|---|
| 9:00–9:30 | **Coffee / Registration** | | |
| 9:30–10:40 | ML | **Cameron Buckner**<br>Houston | Moderate Empiricism and Machine Learning |
| 10:40–11:00 | **Coffee Break** | | |
| 11:00–12:10 | ML | **Rosa Cao**<br>Stanford | Are (Apparently) Successful DNN Models Also Genuinely Explanatory? |
| 12:10–1:20 | **Lunch Break** | | |
| 1:20–3:00 | **Symposium: Representation in Deep Learning Systems** | | |
| 1:20–1:45 | ST | **Fintan Mallory**<br>Oslo | Teleosemantics for Neural Word Embeddings |
| 1:45–2:10 | ST | **Jacqueline Harding**<br>Stanford | Do Probes in NLP Discover Representations? |
| 2:10–2:35 | ST | **Anders Søgaard**<br>Copenhagen | How Language Models View Things |
| 2:35–3:00 | ST | **Tony Chen**<br>MIT | Do Neural Networks Have Concepts? |
| 3:00–4:15 | **Poster Session** | | |
| | PP | **Atoosa Kasirzadeh**<br>University of Edinburgh | Do Large Language Models Understand Linguistic Meaning? |
| | PP | **Wai Keen Vong**<br>New York University | Grounded Language Acquisition Through the Eyes and Ears of a Single Child |
| | PP | **Sreejan Kumar**<br>Princeton University | Characterizing Abstraction Across Natural and Artificial Intelligence |
| | PP | **Will Merrill**<br>New York University | Entailment Semantics Can Be Extracted From an Ideal Language Model |
| | PP | **Julia Minarik**<br>University of Toronto | The Imaginative Shortcomings of Text-to-Image Generators |
| | PP | **Jared Moore**<br>University of Washington | Language Models Understand Us, Poorly |
| | PP | **Nedah Nemati**<br>Columbia University | Exploiting LLMs to Better Understand Assumptions in Social Science Methodologies |
| | PP | **Emin Orhan**<br>New York University | How Much Human-Like Visual Experience Do Current Self-Supervised Learning Algorithms Need in Order to Achieve Human-Level Object Recognition? |
| | PP | **Stephan Pohl**<br>New York University | The Information Gained by Processing a Signal |

| | | | |
|---|---|---|---|
| | PP | **Hokyung Sung**<br>MIT | Predictive Models Are Not Enough for Human-like Cognition A Case Study from Developmental Psychology |
| | PP | **Justin Tiehen**<br>University of Puget Sound | Passing Pearl's Mini-Turing Test |
| 4:15–6:15 | **Panel: What Can Deep Learning Do for Cognitive Science and Vice Versa?** | | |
| 4:15–4:25 | PC | **Ishita Dasgupta**<br>DeepMind | What can we learn from similarities between language model behavior and human behavior? |
| 4:25–4:35 | PC | **Nikolaus Kriegeskorte**<br>Columbia | Neural Network Models as Mechanistic Explanations of Brain Information Processing |
| 4:35–4:45 | PC | **Tal Linzen**<br>NYU/Google AI | What, if Anything, Can Large Language Models Teach Us About Human Language Acquisition? |
| 4:45–4:55 | PC | **Robert Long**<br>Center for AI Safety | Why Cognitive Science Does Not Help AI Progress |
| 4:55–5:05 | PC | **Ida Momennejad**<br>Microsoft Research | A Rubric for Human-like Agents and NeuroAI |
| 5:05–6:15 | **General Panel Discussion** | | |

# Sunday, March 26th – Conference Day 2

**19 West 4th Street, Room 101**

ML: Main Lecture; ST: Symposium Talk.

| | | | |
|---|---|---|---|
| 9:30–10:00 | **Coffee** | | |
| 10:00–11:10 | ML | **Nick Shea**<br>London | The Importance of Logical Reasoning and Its Emergence in Deep Neural Networks |
| 11:10–11:30 | **Coffee Break** | | |
| 11:00–12:10 | ML | **Raphaël Millière**<br>Columbia | Compositionality in Deep Neural Networks |
| 12:40–2:10 | **Lunch Break** | | |
| 2:10–3:20 | ML | **Grace Lindsay**<br>NYU | Developing Neural Systems Understanding |
| 3:20–4:00 | **Coffee Break** | | |
| 4:00–5:40 | **Symposium: Linguistic and Cognitive Capacities of Large Language Models** | | |
| 4:00–4:25 | ST | **Anna Ivanova**<br>MIT | Dissociating Language and Thought in Large Language Models: A Cognitive Perspective |
| 4:25–4:50 | ST | **Nuhu Osman Attah**<br>Pittsburgh | Do Language Models Lack Communicative Intentions? |
| 4:50–5:15 | ST | **Patrick Butlin**<br>Oxford | Functions, Content and Understanding in Large Language Models |
| 5:15–5:40 | ST | **Philippe Verreault-Julien**<br>Eindhoven | Five Lessons Large Language Models Teach Us About Understanding |

# List of Abstracts – Talks

## Saturday, March 25th

### Moderate Empiricism and Machine Learning

***Cameron Buckner***

ML

University of Houston

In this talk, I outline a framework for thinking about foundational philosophical questions in deep learning as artificial intelligence. Specifically, my framework links deep learning's research agenda to classical empiricist philosophy of mind. In recent assessments of deep learning's current capabilities and future potential, prominent scientists have cited historical figures from the perennial philosophical debate between nativism and empiricism, which primarily concerns the origins of abstract knowledge. However, I argue, this debate has often not been invoked in the most useful way. Critics of deep learning frequently paint it as beholden to a radical form of empiricism found in the psychological behaviorists like John Watson and B. F. Skinner, but most research in deep learning fits better with a more moderate and historically grounded form of empiricism found in major philosophical figures like John Locke, David Hume, and William James. This strain of moderate empiricism has not been systematically articulated and defended in the computer science it has inspired. I rebut the radical caricature by explicating the more moderate form, which can be reverse engineered from headline achievements and extracted from the position papers of deep learning pioneers. Moderate empiricism's unifying thread is a commitment to what I call a Domain-General Modular Architecture (a "new empiricist DoGMA") as the best hope of modeling rational cognition in neural-network-based systems. What is missing from the radical caricature—but highlighted by both paradigm historical empiricism and mainstream deep learning—is the critical role played by interactions amongst active, general-purpose faculties (realizing aspects of perception, memory, imagination, attention, and empathy) in the conversion of specific input data into generalizable abstractions. I illustrate the utility of this interdisciplinary connection by showing how it can provide benefits to both philosophy and computer science: computer scientists can continue to mine the history of philosophy for ideas and aspirational targets to hit on the way to more robustly rational artificial agents, and philosophers can see how some of the historical empiricists' most ambitious speculations can be realized in specific computational systems.

## Are (Apparently) Successful DNN Models Also Genuinely Explanatory?

***Rosa Cao***

ML

Stanford University

Do language models and other deep neural networks genuinely capture human capacities or do they merely superficially mimic human behavior? One natural approach is to ask whether these models have internal representations that correspond to the same kinds of internal representations that humans have, playing similar functional roles. It may be useful to attribute contents to internal activations of neural networks in the same ways as we do in biological creatures. But those methods of attribution even in humans involve assumptions and explanatory choices that are no less scientifically and philosophically contentious than the original question – giving us reason to think that the original distinction is not so clear-cut. The original question also echoes earlier debates about instrumentalism vs. realism about scientific theory. In both cases, I suspect that much of what we care about can be understood in terms of differences of degree, rather than a sharp dichotomy. And so rather than asking whether these models truly understand the world, or whether their outputs really have meaning, we might instead ask what aspects of a target phenomenon (whether it be modeling the world or using language) they capture, and to what degree, and under what assumptions.

## Teleosemantics for Neural Word Embeddings

***Fintan Mallory***

IS

University of Oslo

A theory of the representational content for a system should satisfy two criteria at a minimum. It should explain the role that content plays in our lower-level explanations of the system's activity and constrain the contents we ascribe to the system. Teleosemantics is one of the most promising naturalistic theories of what makes something a representation. What it can provide is a relatively precise and general account of how artificial neural networks may develop representations and a means of determining whether these representations are 'intentional' in a theoretically precise sense. Importantly, it gives us a means of distinguishing 'original' from merely 'imputed' intentionality. This talk begins the project of applying teleosemantics to neural language models at the base, Mikolov's Word2Vec algorithm (Mikolov, 2013). Word2Vec was one the first widely-applied method for the production of dense word embeddings. While the main focus of the talk will be on this simple case, a second aim is to indicate how teleosemantics can be applied to artificial neural networks more generally and so I will engage with debates about the role of intentions in determining the function of artefacts and how to individuate vehicles of representation in artificial neural networks.

## Do Probes in NLP Discover Representations?

***Jacqueline Harding***

IS

Stanford University

This talk is concerned with the question: what does it mean for a component of a neural language model to represent a property of an input? I begin with three plausible criteria for assessing representational claims about components of models. First, a component should bear information about the property. Second, the information should be used by the model to perform a task. Third, it should be possible for the component to misrepresent the property. I propose an operationalisation of these criteria using "probing" experiments, in which a supervised classifier is trained to predict the property from the model's intermediate activations. I suggest that the right way to understand both "use" of the information and "misrepresentation" of the property is via appropriate interventions on the model's activations (i.e. on the inputs to probes). I discuss ways of approximating these interventions using ideas from the NLP interpretability literature.

## How Language Models View Things

***Anders Søgaard***

ML

University of Copenhagen

Large-scale pretrained language models (LMs) are said to "lack the ability to connect [their] utterances to the world" (Bender and Koller, 2020). If so, we would expect LM representations to be unrelated to representations in computer vision models. To investigate this, we present an empirical evaluation across three different LMs (BERT, GPT2, and OPT) and two vision encoders (ResNet and SegFormer). Our experiments show that LMs converge towards representations that are isomorphic to those of computer vision models, with dispersion and polysemy both factoring into the alignability of vision and language spaces. This indicates that grounding, as such, possibly, is not a prerequisite for the acquisition of conceptual lexical semantics. One objection to this conclusion, e.g., raised by Yoav Goldberg, is that LMs are trained on data that includes grounding instructions, e.g., code and tables. Such grounding instructions are not restricted to code and tables, however, but are also implicit in day-to-day language use. The role of grounding instructions over distributional evidence, but word embeddings relying on more abstract distributional evidence suggest grounding instructions play a relatively minor role, if any, in the acquisition of conceptual lexical semantics.

### Do Neural Networks Have Concepts?

***Tony Chen***

ML

MIT

Concepts form the cornerstone of human thoughts; any AI system striving for human-level intelligence should therefore possess concepts. With the excitement surrounding neural networks, it is important to ask: how close are current neural networks to having a human-level conceptual system? And even before that, how would we know? Here we attempt to provide a first pass at these questions with the goal of facilitating discussion between AI, cognitive science, and philosophy researchers. Given substantial disagreement on the definition of a concept, we do not attempt to provide yet another theory. Instead, we begin with an incomplete working theory of concepts in neural networks, namely the view of concepts as manifolds in a neural space. We illustrate several ways in which this theory fails to encapsulate psychologically and philosophically rich properties that characterize human-level concepts. Concretely, we (1) provide a non-exhaustive list of properties that a theory of conceptual systems should account for and (2) suggest how some of these properties might be formalized as part of a theory applicable to neural systems. In all, our work aims to lay a common foundation on which researchers from different fields can jointly investigate the nature of concepts in deep neural networks.

## Sunday, March 26th – Conference Day 2

### The Importance of Logical Reasoning and Its Emergence in Deep Neural Networks

***Nicholas Shea***

ML

Institute of Philosophy, University of London

Do deep neural networks reason logically with the representations that emerge in their hidden layers? Does a large language model, after chain-of-thought prompting, reason logically with its outputs? The capacity for broadly logical reasoning is clearly useful (as is variable binding). This paper shows why, when the aim is to produce a general intelligence that draws on information from a wide range of domains, it may be particularly important whether DNNs manage to achieve, or can be engineered to achieve, the capacity for broadly logical reasoning.

## Compositionality in Deep Neural Networks

***Raphaël Millière***

ML

Columbia University

Competent language users can understand and produce a potentially infinite number of novel, well-formed linguistic expressions by dynamically recombining known elements. This is generally taken to support the claim that humans process linguistic expressions compositionally, such that the meaning of complex expressions is determined by the meaning of its constituents, and the way in which they are syntactically combined. Computation over compositionally structured representations has been conjectured to be central not only to linguistic processing, but also to cognition more broadly. Such capacity can be readily accounted for in a classical system that combines discrete symbolic representations into complex representations with constituent structure. By contrast, it has been argued that connectionist systems that do not merely implement a classical architecture lack representations with constituent structure, and are therefore inadequate models of linguistic processing and human cognition. The recent and rapid progress of artificial neural network architectures, ushered by the coming of age of deep learning within the past decade, warrants a reassessment of old debates about compositionality in connectionist models. Deep neural networks called language models, trained on large amounts of text without built-in linguistic priors, have vastly exceeded expectations in many areas of natural language processing. Here, I argue that language models are capable of processing their inputs compositionally, by following systematic rules induced during training instead of shallow heuristics. Accordingly, they encode linguistic information into a structured representational format, even though they fall short of implementing a classical architecture. Specifically, instead of concatenating discrete symbolic representations through strict (algebraic) variable binding, I argue that they can compose distributed (vector-based) representations through a form of fuzzy variable binding enabled by attention mechanisms in the Transformer architecture. I offer both theoretical and empirical support for this hypothesis, and suggest that it goes a long way towards explaining the remarkable performance of language models. The upshot of this analysis is threefold. First, we need not see language models as uninterpretable black boxes. By unraveling the repertoire of computations they induce during training, we can start bridging the gap between behavioral evidence about their performance and claims about their underlying competence. Second, the classicist approach to compositionality is not the only game in town to explain the systematicity of linguistic processing and cognition. Connectionist models need not implement a classical architecture with strict variable binding over discrete constituents to process structured representations compositionally. Third, this non-classical approach to compositionality has a number of characteristics that makes it increasingly attractive not just as an engineering project, but also as an empirically plausible model of linguistic processing in humans. I conclude by offering some reflections about future directions to investigate this claim, and how this line of research may influence cognitive science more broadly.

## Developing Neural Systems Understanding

***Grace Lindsay***

ML

New York University

The emerging field of Interpretable AI aims to understand how trained neural networks work and use that understanding to control their behavior. For over a century, neuroscientists have tried to do the same for the brain. Do these two fields have anything to offer each other? I will argue that they do, and that developing a joint approach to "neural systems understanding" will speed progress in AI and neuroscience. I will describe a research plan to test methods from neuroscience on artificial neural networks and argue that a new vocabulary of concepts will be needed to make progress in understanding these systems. I will also discuss the assumptions and questions raised by this approach, such as: is the algorithmic level the right target for explanation?; what needs to be true to argue that two systems can be submitted to the same analysis?; is understanding required for control and does control demonstrate understanding?; and what sort of concepts will be needed to compactly describe principles of distributed information processing?

## Dissociating Language and Thought in Large Language Models: A Cognitive Perspective

***Anna Ivanova***

ML

MIT

Today's large language models (LLMs) routinely generate coherent, grammatical and seemingly meaningful paragraphs of text. This achievement has led to speculation that these models are—or will soon become—"thinking machines", capable of performing tasks that require knowledge and reasoning. In this talk, I will argue that, when evaluating LLMs, we should distinguish between their formal linguistic competence—knowledge of linguistic rules and patterns—and functional linguistic competence—understanding and using language in the world. This distinction stems from modern neuroscience research, which shows that these skills recruit different mechanisms in the human brain. I will show that, although LLMs are close to mastering formal linguistic competence, they still fail at many functional competence tasks, which in humans require drawing on various non-linguistic skills. I will conclude by discussing the implications of the formal/functional competence distinction for training and evaluating future AI models.

## Do Language Models Lack Communicative Intentions?

***Nuhu Osman Attah***

ML

University of Pittsburgh

In recent work, some psychologists/linguists, AI researchers, and philosophers (e.g., Shanahan 2022, Bender & Koller 2020, Montemayor 2021, Bender et al. 2021) have argued that language models do not possess genuine linguistic competence on the ground that they lack communicative intention. In this presentation I argue that (even barring the untenability of the strong Gricean assumption about the nature of language which ostensibly underlies this position) some of the key arguments used to defend this conclusion are severely flawed. Even though the general conclusion (that language models do not possess genuine linguistic competence) might be independently true, the "communicative intention argument" fails to demonstrate that it is.

## Functions, Content and Understanding in Large Language Models

***Patrick Butlin***

ML

University of Oxford

In previous work, I have argued that LLMs cannot understand human utterances. One argument for this claim is that they lack sensory grounding. However, my argument was different: I claimed that LLMs perform only purely linguistic tasks, that representational content is determined by function, and that LLMs are therefore only able to form representations with content which concerns language itself. An LLM cannot understand the word 'skin' because the properties of skin itself – as opposed to the word 'skin' – are not relevant for its task. I am now less confident that this argument is correct. One objection is that information about the non-linguistic world can be useful even for purely linguistic tasks. A second is that fine-tuning methods such as RLHF make LLM functions less clear. A third is that there are many entities which humans can think about even though the only tasks we perform with respect to them are linguistic. These objections do not show that the argument fails, but do show that the broad philosophical principles it appeals to are not sufficient to settle the issue. A more detailed analysis is needed.

## Five Lessons Large Language Models Teach Us About Understanding

***Philippe Verreault-Julien***

ML

Eindhoven University of Technology

This paper explores the implications of state-of-the-art large language models (LLMs) such as GPT-4 for the concept of understanding, both in philosophy and natural language processing (NLP) practice. LLMs challenge the conventional notion that language models lack understanding. However, what constitutes 'understanding' is contentious in both philosophy and NLP. The paper identifies three lessons for philosophers and two for NLP practitioners from studying LLMs. For philosophers, the paper suggests that unpacking the nature of 'grasping' and exploring other abilities like abstraction and analogy may be crucial to understanding. Furthermore, it argues that LLMs put pressure on accounts that view understanding as being compatible with luck or falsehood. For practitioners, the paper argues that the capabilities of LLMs demonstrate that understanding comes in degrees and that improving LLMs' understanding may require exploiting and representing other information besides statistical correlations. Overall, the paper suggests that LLMs provide an exciting opportunity for research at the intersection of philosophy and NLP.

# List of Posters

## Can Large Language Models Understand Meaning?

***Atoosa Kasirzadeh***

University of Edinburgh

This paper investigates the question, "Can large language models understand meaning?" To address this, we first need to establish a comprehensive characterization of "meaning." I delve into the theories of meaning literature and distinguish between two types of meaning questions: (i) semantic questions, which ask, "What is the meaning of a particular linguistic expression?" and (ii) foundational questions, which ask, "What mental or social facts about a person, group, or society give linguistic expressions the meaning they possess?" I argue that while large language models are making progress in capturing semantic meaning through mechanisms of embeddings, self-attention, and multi-headed attention, they are unable to grasp foundational meaning on their own. To understand meaning requires capturing both semantic and foundational aspects. I close by some constructive reflections on this argument.

## Grounded Language Acquisition Through the Eyes and Ears of a Single Child

***Wai Keen Vong***

New York University

Starting around 6-9 months of age, children begin acquiring their first words, learning to ground linguistic symbols to their visual counterparts. How much of this knowledge about grounded word meanings is learnable from sensory inputs and relatively generic learning mechanisms, and how much requires stronger inductive biases? Using a dataset of longitudinal head-mounted camera recordings from a single developing child aged 6 to 25 months, we trained a multimodal neural network on correlated visual-linguistic data streams and examined the knowledge it acquired. We find that our model can acquire many word-referent mappings present in the child's everyday experience from tens of noisy examples, learning multimodal representations that enable zero-shot generalization to highly novel visual referents and aligning its visual and linguistic conceptual systems. These results demonstrate that critical aspects of grounded word meaning are learnable from a subset of a single child's sensory input using generic multimodal learning mechanisms.

## Characterizing Abstraction Across Natural and Artificial Intelligence

***Sreejan Kumar***

Princeton University

Humans have always been motivated to make sense out of the confusing world we live in. According to early psychological theory, a fundamental aspect of human general intelligence is the presence of strong inductive biases that capture the abstract structure of the world and enable effective generalization beyond specific learning contexts. In this work, we develop a common task paradigm to compare inductive biases towards abstraction in humans, deep reinforcement learning agents, and non-human primates. We use a controlled stimulus space of two-dimensional grids and use large-scale behavior studies to sample the space of abstract concepts humans associate with this stimulus space. We then formulate a meta-reinforcement learning task paradigm where the task distribution directly samples from this space of abstractions. The results show that humans generalize better to test tasks that contain abstract concepts, while deep reinforcement learning agents and non-human primates generalize better to control tasks devoid of abstractions. Additionally, co-training the artificial agent with representations from human-written language descriptions of the stimuli or symbolic programs that draw the stimuli guides it during training to exhibit human-like generalization. Our results suggest that natural language and domain-specific symbols contain useful abstract knowledge necessary to emerge human-like intelligence.

## Entailment Semantics Can Be Extracted from an Ideal Language Model

***Will Merrill***

New York University

Language models are often trained on text alone, without additional grounding. There is debate as to how much of natural language semantics can be inferred from such a procedure. We prove that entailment judgments between sentences can be extracted from an ideal language model that has perfectly learned its target distribution, assuming the training sentences are generated by Gricean agents, i.e., agents who follow fundamental principles of communication from the linguistic theory of pragmatics. We also show entailment judgments can be decoded from the predictions of a language model trained on such Gricean data. Our results reveal a pathway for understanding the semantic information encoded in unlabeled linguistic data and a potential framework for extracting semantics from language models.

## The Imaginative Shortcomings of Text-to-Image Generators

***Julia Minarik***

University of Toronto

Generative AI like DALLE-2 (D2), Midjourney, and Stable Diffusion are artificial intelligences that generate artworks conditional on short, descriptive, natural language prompts. My aim is to interrogate their imaginative capacities. Following a suggestion by David Holz of Midjourney, I see these image generators as prosthetic imaginations: AI that extend the human imagination by allowing people to bring imagined images outside of their head and calcify them. These machines are impressively imaginative: they create beautiful artworks without explicit human guidance. That said, there are multiple points of translation that open the door to error: (i) the human must accurately translate their imagined content into a descriptive prompt; (ii) the TIG must understand that prompt by translating it into a text embedding and relate it to an image embedding; and (iii) the diffusion model (part of the TIG) must accurately generate an appealing image guided by the embeddings. I argue that these machines face fundamental imaginative challenges in each of the steps of the translation process. The most interesting of these is that TIGs interpret descriptive prompts literally, meaning that they cannot imagine – and therefore generate – metaphorical representations. This means that generative AI lack a key ability of the human artistic imagination.

## Language Models Understand Us, Poorly

***Jared Moore***

University of Washington

TBC.

## Exploiting LLMs to Better Understand Assumptions in Social Science Methodologies

***Nedah Nemati***

Columbia University

What do we stand to learn from 'silicon sampling' with large language models (LLMs) — that is, using LLMs to 'simulate' human beings in social science? Recent work by Argyle et al. (2023) claims that this allows researchers to use LLMs to "advance understanding of humans and society." Instead, I claim silicon sampling helps us gain better knowledge of social scientific methods and how to test with them. Specifically, one overlooked value of LLMs consists in how it may serve as a tool for philosophy of science. In the case of silicon sampling, I show how this relates to sampling methods constructing human data patterns rather than descriptively identifying them. Along the way, I will discuss three ethical issues surrounding silicon sampling. First, there is a tradeoff between the accuracy of LLMs used for silicon sampling and ethical concerns about training LLMs with harmful information. Second, using silicon sampling opens the door to developing increasingly refined and powerful tactics of persuasion, which may exacerbate political power asymmetries and spread misinformation. Third, the idea that LLMs represent deep features of the human essence can appear illicit when there is only a focus on the matching between such models and their human counterparts.

## How Much Human-Like Visual Experience Do Current Self-Supervised Learning Algorithms Need in Order to Achieve Human-Level Object Recognition?

***Emin Orhan***

New York University

TBC

## The Information Gained by Processing a Signal

***Stephan Pohl***

New York University

Intuitively, processing a signal can extract information about features from the signal. The output of an image classifier network, in some sense, carries more information about image categories than the unprocessed image. Late layers in a network carry more information about complex features than early layers. What is the information gained by processing a signal? We need to distinguish two dimensions of the information a signal carries about a feature. First, the feature information is the mutual information between the signal and the feature. By the data processing inequality, the feature information cannot be increased by the processing of a signal. Secondly, the model uncertainty is the information needed in order to specify a model that decodes the feature from the signal. The information gained by processing a signal is the reduction in model uncertainty. I train an image classifier on a synthetic dataset and show that the minimum description length method offers the best measures of feature information and model uncertainty across the layers of the image classifier. Related measures like the information decodable under computational restrictions or measures of the information bottleneck method fail to track model uncertainty.

## Predictive Models Are Not Enough for Human-like Cognition A Case Study from Developmental Psychology

***Hokyung Sung***

MIT

Recent work from Perez and Feigenson (2022) showed that infants are triggered to play with a stimulus if it was expectation-violating (i.e. surprising), but not when an explanation for that violation was shown immediately afterwards. Here, I argue that this finding challenges extant approaches towards building human-like world models and intuitive physics engines. First, I claim that current approaches to curiosity-based exploration within the field of deep reinforcement learning which utilize predictive (forward-in-time) world model architectures cannot reproduce this pattern of exploration behavior. While existing prediction error-based accounts may capture how infants were initially led to explore more when faced with surprising stimuli, such models are insufficient for reproducing the reduced level of curiosity after presentation of the "explanation." This is because there is no mechanism with which a particular explanatory piece of information can "resolve" prediction error of a past state. Secondly, I argue that a characterization of the intuitive physics engine in the brain as simulating the future progression of physical dynamics is also insufficient to account for this behavior, for analogous reasons. I conclude that a non-predictive formulation of world models and intuitive physics engines is necessary for a behaviorally faithful account of infant-like cognition.

## Passing Pearl's Mini-Turing Test

***Justin Tiehen***

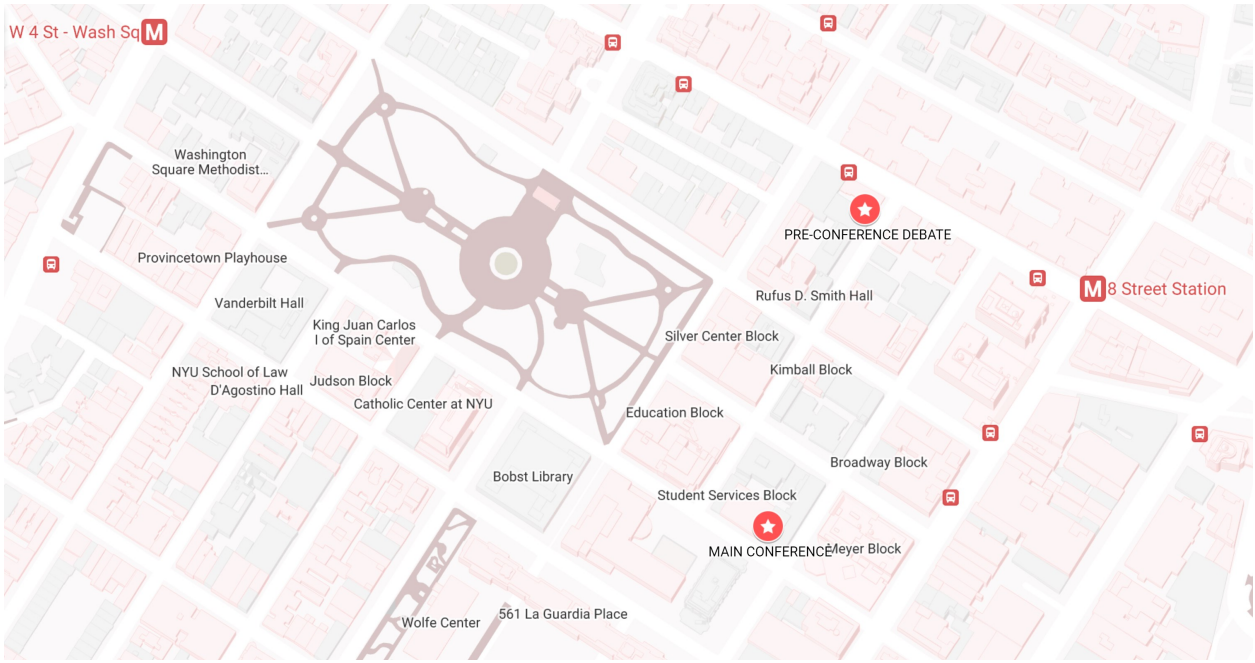
University of Puget Sound

Causation is often thought to pose an especially serious challenge for deep learning models in AI. In The Book of Why, Judea Pearl proposes a "mini-Turing test" where a machine is presented with a simple story and asked questions specifically about causation. Pearl predicts that typical deep learning systems will be unable to pass the test. In his recent New York Times piece, Noam Chomsky advances a similar claim about ChatGPT. In fact, though, ChatGPT does fairly well answering causal/counterfactual questions. Since this is so, I revisit Pearl's argument, which I understand as turning on an underdetermination thesis about his Causal Ladder: given a causal model, the truths at the lowest (associational) rung of the Ladder do not entail the truths at the higher (interventional and counterfactual) rungs. I argue that even granting such underdetermination, an LLM operating entirely at the lowest rung could potentially pass the mini-test by taking advantage of the point that although causal/counterfactual properties (worldly entities) are unobservable, causal/counterfactual words (linguistic entities) are fully observable, and so can be "seen" in the training data. But I then go on to suggest this provides a reason to be skeptical about the adequacy of Pearl's mini-Turing test.

# Useful Information

The **pre-conference debate** on Friday, March 24th will be held at the **Cantor Film Center**, Room 200, 36 East 8th Street.

The **main conference** on Saturday and Sunday, March 25-26th will be held at **19 West 4th Street**, Room 101.



# Sponsors

The Philosophy of Deep Learning conference is jointly sponsored by the Center for Science and Society at Columbia University and the Center for Mind, Brain, and Consciousness at New York University.

COLUMBIA UNIVERSITY
Presidential Scholars in Society and Neuroscience

NYU CENTER FOR MIND, BRAIN, AND CONSCIOUSNESS