Philosophy and AI
Lecture 9-10: Philosophy of Language

Marco Degano

11-12 March 2025

Readings

Required:

- Glymour 2015: chapter 15, pp. 298 – 307
  - Required: What are Meanings?; Truth Conditions; Meaning as Use; The Private-Language Argument
- Millière, R., & Buckner, C. (2024). `https://arxiv.org/abs/2401.03910`
  - Required: Intro section 3; 3.1 Compositionality; 3.3 Language understanding and grounding; 3.4 World models

Optional:

- Lepore, E., & Smith, B. C. (2006). The Oxford Handbook to the Philosophy of Language. `https://academic.oup.com/edited-volume/38631`
- John Searle, Introduction to Philosophy of Language, Lectures recordings `https://youtu.be/Uk5pIzCNOzU?si=nAhT82F0hd0k1_Il`

Outline

Outline

Philosophy of language

- Philosophy of language investigates the nature of language.
- It examines how language relates to reality, thought, and understanding.
- Key questions (relevant to AI):
  - What does it mean for words (or tokens) to have meaning?
  - How do symbols refer to objects in the world?
  - What is the relationship between language and thought?
  - How does understanding emerge from symbol manipulation?
  - How is language grounded in communication and social contexts?

Meaning

- Fundamental question: What is the meaning of a word?
- Consider the sentence ‘The cat is on the mat’:
  - What exactly is the meaning of ‘the cat’?
  - How does the phrase ‘the cat’ connect to reality?

Meaning and Reference

- **Referential Theory**: Words mean what they refer to
  - 'The cat' refers to a particular animal with certain characteristics
  - Meaning = reference relationship to world objects
- Historical roots:
  - Traces to Aristotle's view of words as signs for things.
  - Formalized in modern analytic philosophy.
  - Aligns with how we learn language as children (pointing at a cat and saying "cat").

Puzzles for the Referential Theory

- Two fundamental problems for a purely referentialist view:
- **1. Problem of Non-referring terms**:
  - Example: 'Phlogiston was thought to be the cause of combustion.'
  - Phlogiston: a hypothetical substance that does not exist
  - Question: What is the meaning of a word whose referent does not exist?
  - If meaning = reference, these terms would be meaningless (but they're not!)
- **2. Problem of Informative Identity:**
  - Compare: 'Amsterdam is the capital of the Netherlands' vs. 'Amsterdam is Amsterdam'.
  - Both statements involve terms with identical referents.
  - Yet the first might be informative while the second is trivial.
  - If meaning = reference, these should convey the same information (but they don't!).

Sense and Reference: Frege's Solution

Gottlob Frege (1848–1925) distinguished between:

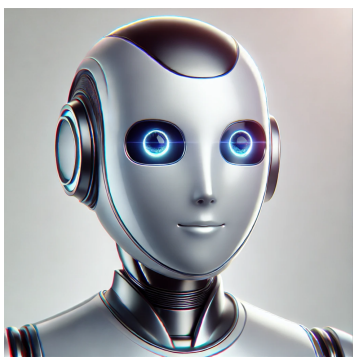
- Sinn (sense): The mode of presentation or conceptual content
- Bedeutung (reference): The actual object designated in the world

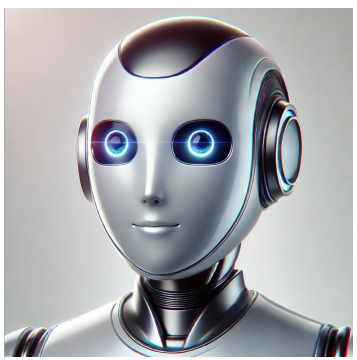
Frege's famous example:

- *Hesperus* (the "evening star") and *Phosphorus* (the "morning star") both refer to the same celestial body, Venus.
- However, they have different senses because each phrase presents Venus under a different characterization (evening vs. morning).

Echo vs. Nova

- Imagine two chatbots, "Echo" and "Nova", widely known for different capabilities.
- Echo is famous for answering math questions.
- Nova is renowned for creative advice.
- Unknown to users, Echo and Nova are the same underlying AI model.





- **Analysis:**
  - The referent is identical (the same AI model)
  - The senses are different (the math expert vs. the creative advisor)

How Sense Determines Reference

- For Frege, sense is the mechanism by which an expression picks out its reference:

  "The sense of an expression is that wherein the mode of presentation is contained, and it determines the reference."

- Different senses can lead to the same referent
  - "Mark Twain" vs. "Samuel Clemens" (same person, different senses)
  - "Echo" vs. "Nova" (same model, different presentations)
  - "$3 + 1$" vs. "2×2" (both refer to 4, different computational paths)
  - "localhost" vs. "127.0.0.1" (same network location, different notations)

Sense and Reference in Word Embeddings

- Can vector representations in NLP capture Frege's distinction between sense and reference?
- Cosine similarities in `word2vec` for co-referential terms:
  - 'Hesperus' and 'Phosphorus': 0.9502
  - 'Zeus' and 'Jupiter': 0.7410
  - 'Cicero' and 'Tully': 0.7286
  - 'Istanbul' and 'Byzantium': 0.6860
  - 'Sodium' and 'Natrium': 0.6742
  - 'water' and 'h2o': 0.6227
- **Questions for you:**
  - What kind of architecture could fully capture the sense/reference distinction?
  - Could embedding spaces have specialized subspaces for sense vs. reference?
  - Are multi-layered transformers better at distinguishing sense from reference?

Sense and Reference for Sentences

- Extending Frege's distinction to entire sentences:
  - The **reference** of a sentence is its truth value (True/False).
  - The **sense** of a sentence is the thought or conceptual content expressed.
  - As with names, the thought/sense of a sentence determines its reference/truth value.
- **Example:**
  - 'Echo is a smart chatbot' and 'Nova is a smart chatbot'.
  - Same reference (both are true statements).
  - Different senses (different grounds for truth judgment).
  - I judge 'Echo is a smart chatbot' as true, because Echo can answer math questions. While I judge 'Nova is a smart chatbot' as true, because Nova can be creative.

Compositionality

- **Principle of Compositionality**: The meaning of a complex expression is determined by:
  - The meanings of its constituent parts
  - The rules used to combine those parts
- **Sense compositionality:**
  - The sense (thought) of a sentence is built from the senses of its terms
  - "The cat is on the mat" = sense("the cat") + sense("is on") + sense("the mat")
  - Result: a structured thought about a cat on a mat.
- **Reference compositionality:**
  - The reference (truth value) depends on references of parts and syntax
  - "The cat is on the mat" is True if and only if:
  - The referent of"the cat" stands in the relation denoted by"is on" to the referent of"the mat".

Intersubstitutivity

- **Principle of Intersubstitutivity:**
  - Expressions with the same referent should be interchangeable while preserving truth.
  - Example: "Venus is bright" $\equiv$ "The Morning Star is bright" $\equiv$ "The Evening Star is bright"
  - All three statements have the same truth value/reference (True)
  - But they might have a different thought/sense.

How Frege's Theory Solves the Puzzles

- **Problem 1: Non-referring terms**
  - Frege's solution: Terms like "Pegasus" and "Phlogiston" have a sense but no referent.
  - Sense provides meaning even without a corresponding object.
  - Statements containing such terms lack truth value (reference) but retain their sense.
- **Problem 2: Informative identity statements**
  - "Echo = Echo" vs. "Echo = Nova"
  - Same reference (both statements are true)
  - Different senses (different cognitive content)
  - The latter is informative because it connects two distinct modes of presentation

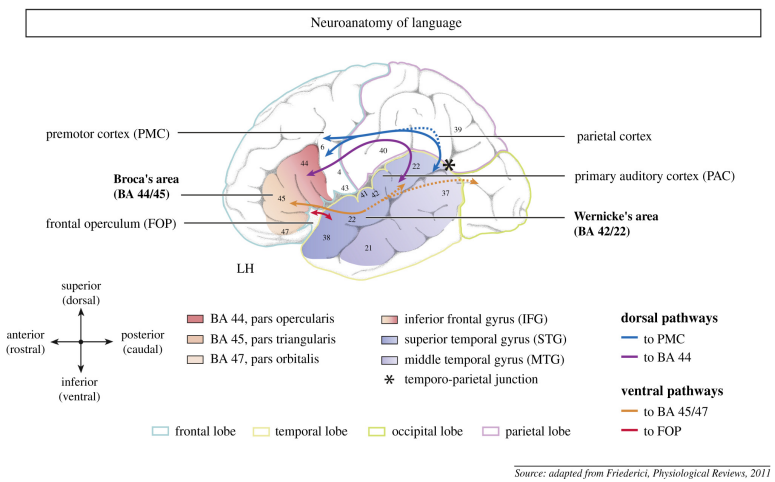
Outline

Why Compositionality Matters

- The principle of compositionality explains crucial features of language and thought:
- **Productivity:**
  - Humans generate and understand infinite novel expressions from finite vocabulary.
  - Example progression: "The cat" $\rightarrow$ "The cat on the mat" $\rightarrow$ "The cat on the mat that belongs to the woman" $\rightarrow$ ...
  - This infinity requires systematic composition, not memorization.
- **Systematicity:**
  - Related expressions are systematically understood together.
  - If someone understands "John loves Mary," they also understand "Mary loves John".
  - Knowledge of semantic roles transfers across different sentences.

Why Compositionality Matters

- **Cognitive/Computational Efficiency:**
  - Reusing building blocks reduces storage and processing demands.
  - Learning general rules is more efficient than memorizing every possible sentence.
- **Empirical evidence:** Neuroimaging shows hierarchical processing of language structure.



Friederici, Angela D. "Hierarchy processing in human neurobiology: how specific is it?." *Philosophical Transactions of the Royal Society B* 375.1789 (2020) https://doi.org/10.1098/rstb.2018.0391

The Classical Critique of Neural Networks

- Fodor & Pylyshyn's (1988) influential critique of artificial neural networks (ANNs):
- **Core argument:** ANNs fundamentally lack the architecture to support compositionality.
  - Human cognition requires a "language of thought" with:
    - Discrete symbols with explicit semantic content.
    - Formal rules for compositional combination.
    - Structure-sensitive operations on representations.
  - ANNs with their distributed representations cannot capture these features.
- **Dilemma posed for neural networks:**
  - Either: ANNs fail to model human-like compositional cognition.
  - Or: ANNs merely implement classical symbolic architectures underneath, with no independent explanatory value.

Language of Thought: An Example

- **Consider two sentences:**
  `(1) John loves Mary` and `(2) Mary loves John.`
- **In a Language of Thought (LoT) format:**
  - Sentence (1) might be internally represented as `LOVES(JOHN,MARY)`.
  - Sentence (2) would be `LOVES(MARY,JOHN)`.
- **Why this matters:**
  - The LoT hypothesis says we have mental "symbols" for `JOHN`, `MARY`, and `LOVES`, and rules specifying how to arrange them.
  - Changing the order of these symbols changes the meaning: *who loves whom*.
  - Shows systematicity: understanding "John loves Mary" implies understanding "Mary loves John," because it's just rearranging symbols in the LoT.

The Connectionist Response

- **Connectionist counterargument:**
  - Distributed representations (continuous vectors) may achieve compositionality differently.
  - Compositionality might emerge from statistical learning without explicit symbols.
  - Neural systems might implement "functional compositionality" without classical structure.
- **Key questions:**
  - Can neural networks model productivity and systematicity?
  - Must compositional behavior rely on classical symbolic structure?
  - What empirical tests could decide between these perspectives?

Reevaluation in the Era of Large Language Models

- The success of transformer-based LLMs has prompted reevaluation of the classical critique:
- **Shift in research focus:**
  - From theoretical arguments to empirical evaluation
  - Central question: Do LLMs demonstrate compositional generalization?
  - Testing whether models can recombine learned elements in novel ways
- **Methodological challenges:**
  - Hard to distinguish memorization from compositional understanding
  - Need for controlled tests of compositional abilities

Testing Compositional Generalization

- **Principles of Compositional Generalization Testing:**
  - **Systematicity:** Testing if models recognize structural patterns across different content.
  - **Productivity:** Evaluating ability to handle novel combinations of known elements.
  - **Distribution Shift:** Train and test distributions must differ in principled ways.
  - **Controlling for Memorization:** Ensuring models aren't simply recalling training instances.

Testing Compositional Generalization

- **The SCAN Dataset (Lake & Baroni, 2018):**
  - **Purpose:** Tests compositional generalization in neural networks
  - **Structure:** Mapping natural language commands to action sequences
  - **Example:** "walk twice" $\rightarrow$ "WALK WALK"
  - **Challenging splits:**
    - Length generalization: Train on short sequences, test on longer ones.
    - Primitive generalization: Hold out one primitive verb in specific contexts.
    - Compositional generalization: Combine known elements in new ways.
- **Other benchmarks:** COGS (Kim & Linzen, 2020), CFQ (Keysers et al., 2020), CLEVR (Johnson et al., 2017)

Evolution of Model Performance on Compositional Tasks

- **Early Neural Network Performance (2018-2020):**
  - **Sequence-to-Sequence Models:** Poor systematic generalization (<20% on SCAN splits)
  - **Convolutional Seq2Seq:** Moderate improvements but still struggling
  - **Basic Transformers:** Significant train-test performance gaps
- **Recent Transformer Advancements (2021-2025):**
  - **Specialized architectures:** Near-perfect SCAN performance
  - **Scale effects:** Larger models show better compositional abilities

Classical Interpretation

- **Symbolic Implementation Hypothesis:**
  - ANNs succeed by implementing a language of thought architecture.
  - Neural activations encode discrete symbolic structures.
  - Neural operations implement compositional rules.
- **Evidence Supporting This View:**
  - In transformer architectures, some attention heads track explicit linguistic dependencies (e.g., heads specialized for subject-verb agreement)
- **Proponents:** Quilty-Dunn et al. (2022), Pavlick (2023), Marcus (2020)

Distributed Processing Interpretation

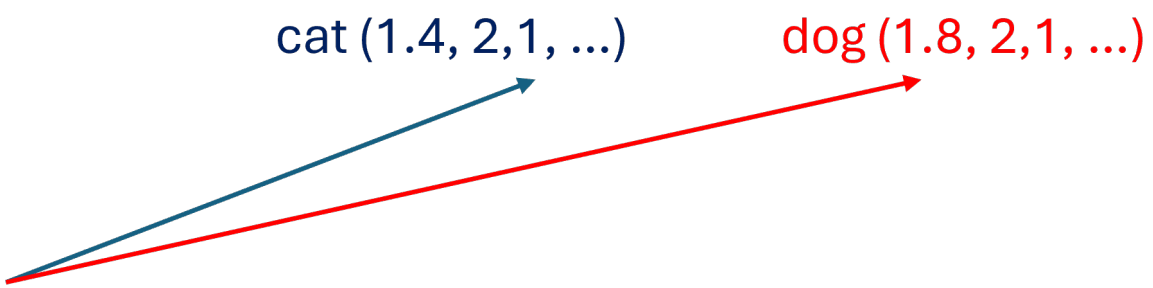
- **Distributed Processing Hypothesis:**
  - ANNs achieve compositionality without implementing classical symbolic systems.
  - Vector representations encode meaning holistically.
  - Compositional behavior emerges from statistical learning.
- **Evidence Supporting This View:**
  - Graceful degradation under noise unlike symbolic systems.
  - Representations show continuous rather than discrete structure.
  - Meaningful intermediate states unlike classical symbol manipulation.
- **Proponents:** McClelland et al. (2010), Rogers & McClelland (2014), Lake & Baroni (2023)

The Revisionist Connectionist Position

- **Classical "Rigid" Assumptions Questioned:**
  - Are discrete symbols necessary for compositionality?
  - Must systematicity be all-or-nothing?
  - Are explicit rules required for structure-sensitive processing?
  - Is algorithmic transparency necessary for explanatory adequacy?
- **"Revisionist" Perspective:**
  - Distributed microfeatures (e.g., patterns of activation across many interconnected elements) represent meaning implicitly.
  - Features emerge through learning rather than explicit design/rules.
  - Processing operates on patterns rather than symbols.
  - Representations are context-sensitive and dynamic.

The Continuity Principle

- **Smolensky (2022)'s Continuity Principle:**
  - Information encoding and processing should use continuous real numbers rather than discrete symbols.
  - Information encoding arise from a high-dimensional space of distributive microfeatures.
- Word embeddings can be considered as distributive microfeatures.
- But the notion of distributive microfeature is more general: any representation where information is spread across multiple dimensions rather than being assigned to a single unit (e.g., tensors, feature maps, hidden states in transformers, . . .).



Three Critical Advantages of Continuity

- **1. Similarity-Based Generalization**
  - Knowledge transfers along similarity gradients in vector space.
  - Word embeddings capture semantic relationships.
  - Generalization to novel but similar examples.
- **2. Tractable Approximation Solutions**
  - Continuous optimization vs. discrete combinatorial explosion.
  - Gradient-based learning through continuous functions.
  - Enables scaling to billions of parameters.
- **3. Representational Optimization**
  - Representations and computations learned simultaneously.
  - Self-attention creates context-dependent representations.
  - Specialized subspaces for different linguistic phenomena.

Microfeatures and Intuitive Understanding

- **The Microfeature Challenge:**
  - ANNs lack discrete constituents. They rely on distributed microfeatures.
  - **What Are Microfeatures?**
    - Tiny, distributed patterns in vectors (e.g., "cat" as a mix of "furry", "whiskers", "meow").
    - Emerge from training, not predefined rules.
    - Overlap and interact in complex, non-intuitive ways.
- **Philosophical Tension:**
  - **Human Experience:** Compositionality feels transparent, as rules govern our language.
  - **AI Reality:** Microfeatures are opaque, success without intuitive clarity.
  - **Core Question:** Can we accept a form of compositionality we cannot intuitively grasp?

Outline

Reference and Descriptions

- Names typically have a reference (what they point to in the world)
- **Frege** suggested that the sense determines the reference.
- **Russell** proposed that reference occurs via descriptions.
- **Descriptivist theory:** The meaning of a term is given by a description (or set of descriptions), which determines its referent
- Example: "Aristotle" means "the philosopher who taught Alexander the Great and wrote Nicomachean Ethics"
- **Problem:** What if Aristotle hadn't taught Alexander? Would "Aristotle" still refer to the same person?
- This reveals a fundamental weakness in pure descriptivism.

Reference and Causal Chain

- **Causal theory of reference**: **Kripke** proposed that names refer via a historical causal chain, not descriptions.
  1. An initial "baptism" fixes the name to an object (e.g., parents name a baby "Aristotle")
  2. The name spreads through a community via communication, linking back to that original event.
  3. Reference persists even if later users know little about the object.
- Names function as rigid designators that refer to the same individual across all possible situations (e.g., also the situation in which Aristotle did not teach Alexander.)
- **Solution:** Even if Aristotle hadn't taught Alexander, "Aristotle" would still refer to him.
- Reference is external and social, not merely mental or descriptive
- This shifts from internal meaning to external causal history

Reference and Reality

- Kripke's causal theory influenced **Putnam's** semantic externalism.
- **Semantic Externalism:** "Meanings ain't in the head" - reference depends on the external environment.
- **Idealism:** Reference is formed by internal mental constructs.
- **Twin Earth Thought Experiment:**
  - Identical twins on Earth and Twin Earth both use "water"
  - On Earth, "water" refers to H_2O; on Twin Earth, to XYZ
  - Despite identical internal mental states, they refer to different substances
  - The referent is determined by the environment, not only internal mental content

AI Challenge: The Symbol Grounding Problem

- **Symbol Grounding Problem** (Harnad): How do artificial systems assign meaning to symbols in a way that is not merely based on other symbols but is instead meaningfully connected to the real world?
- **Human grounding:** Through sensory experience (seeing a cat, feeling its fur) and social interaction.
- **Symbolic AI:** Symbols (e.g., "CAT") are arbitrary tokens requiring external interpretation.
- **LLMs:** Work with vectors in high-dimensional space, but still lack direct grounding.
- The formal challenge: Establish mapping $f : S \to R$ where:
  - S = symbols/vectors in the model
  - R = real-world referents
- Without grounding, do AI systems truly "understand" or merely manipulate symbols?

LLMs and Inherited Reference

- **Mandelkern & Linzen (2023) Hypothesis**:
  - LLMs lack direct interaction with referents but train on human-generated text.
  - These corpora encode causal chains of reference, preserving linguistic conventions without direct experiential grounding.
  - Example: "Tiger" in training data reflects biologists' usage, indirectly linking to *Panthera tigris*.
- **Limitations:**
  - Inherited, not direct: LLMs acquire meaning via patterns in text, but lack the embodied experience that anchors human understanding.
  - Vulnerability to textual biases: Without grounding in reality, LLMs depend entirely on the correctness and completeness of textual data.

Can RLHF Ground LLM Representations?

- **Proposal** (Mollo & Millière, 2023):
  - Reinforcement Learning from Human Feedback (RLHF) may enhance grounding
  - Human feedback serves as an extralinguistic standard.
  - Example: Feedback on "water boils at 100°C" ties output to physical reality.
- **Mechanism:**
  - RLHF introduces a reward signal (e.g., +1 for accuracy, -1 for errors)
  - LLM learns to prioritize outputs reflecting true referents
  - Adds a normative layer beyond pure statistical learning (think also of post-training verification methods)
- **Limitations:**
  - Still no direct sensory access - grounding is mediated by human judgment.
  - Does alignment with human judgment constitute genuine understanding?

Multimodal Models and Grounding

- **Multimodal models** combine language with visual processing.
- **Potential advantage:** Direct connection between linguistic tokens and visual perception.
  - Example: "Cat" token associated with thousands of cat images.
  - May provide partial solution to grounding problem via cross-modal alignment.
- **Limitations:**
  - Still lacks embodied interaction with the world
  - No causal manipulation of objects-only passive perception
  - Question: Is perception without interaction sufficient for grounding?
- **Philosophical question:** Do multimodal representations constitute a form of weak grounding, or merely a more sophisticated form of pattern matching without true understanding?

Outline

Meaning as Use

- Traditional view: The meaning of a sentence is determined by its truth conditions.
- Truth-conditional semantics: A sentence's meaning is the conditions under which it is true or false.
- Late Wittgenstein challenged this view in *Philosophical Investigations* (1953).
- **Meaning as use:** The meaning of a word or sentence is its function in language practices.
- Key insight: Meaning is dynamic, context-dependent, and embedded in human activities.
- Example: "Game" - No single truth-condition unites chess, soccer, and tag, but their uses reveal family resemblances.

Language Games

- **Language games:** Rule-governed activities where words function within specific contexts.
- Examples that challenge truth-conditional semantics:
  - "Pass the salt" - Meaning lies in the request/action, not a truth-value
  - "Checkmate" - Meaningful only within chess, not as a standalone proposition
  - "How are you?" - Social ritual, not primarily an information-seeking query
- Properties of language games:
  - Context-specific: Words function differently across games
  - Socially embedded: Meaning requires shared practices or"forms of life"
- Many meaningful utterances lack truth-values: commands, questions, exclamations
- Wittgenstein: "The meaning of a word is its use in the language" (PI §43)

Family Resemblances

- **Essentialism** (traditional view):
  - Words have fixed definitions with necessary and sufficient conditions
  - Frege/Russell: Meaning tied to logical reference or sense
- Wittgenstein's **rejection of essentialism**:
  - No single trait defines all instances of most concepts
  - Instead: overlapping similarities ("family resemblances")
- Example: "Game"
  - Chess (strategy, competition, rules)
  - Soccer (physical, team-based, competition)
  - Solitaire (solo, rules, pastime)
  - No single feature unites all games
- Meaning emerges from a network of related practices, not fixed definitions

The Private Language Argument

Private Language: A hypothetical language referring only to private sensations, intelligible to one person alone.

1. **Language Requires Rules:**
   - Terms must have criteria for correct application.
   - Example: *Red* follows public standards of color identification.
2. **Rules Require Standards of Correctness:**
   - Must distinguish between correct and incorrect applications.
   - *Thinking one is following a rule is not following a rule* (PI §202)
3. **Private Language Lacks External Standards:**
   - If S refers to a private sensation, what ensures consistent application?
   - No way to verify consistency beyond subjective impression.

World Models and Forms of Life

- Wittgenstein's "forms of life": Shared practices reflecting understanding of the world.
- Language games presuppose practical understanding of:
  - Physical objects and their properties
  - Causal relationships
  - Human intentions and social norms
- Example: "Pass the hammer" assumes:
  - Understanding of tools and their functions
  - Physical grasp of handling objects
  - Social practice of cooperation
- Do LLMs capture the "forms of life" behind language?
- **World models:** internal representations that a system (biological or artificial) uses to understand, predict, and interact with the world.
- Can LLMs learn world models from text alone?

Evidence for World Models in LLMs

- Experimental approach (Wang et al., 2023):
  - Test whether LLMs capture world knowledge beyond text patterns
  - Challenge: Generate executable text-based games from prompts
- Results with GPT-4:
  - 28% runnable games with one-shot learning
  - 57% success rate with self-correction
  - Games require modeling causal relationships (e.g., fire needs fuel, oxygen)
- Interpretation:
  - Success suggests rudimentary simulation of object interactions
  - Limitations: partial success rates; lack of consistency
  - Open question: Are these genuine world models or sophisticated pattern matching?

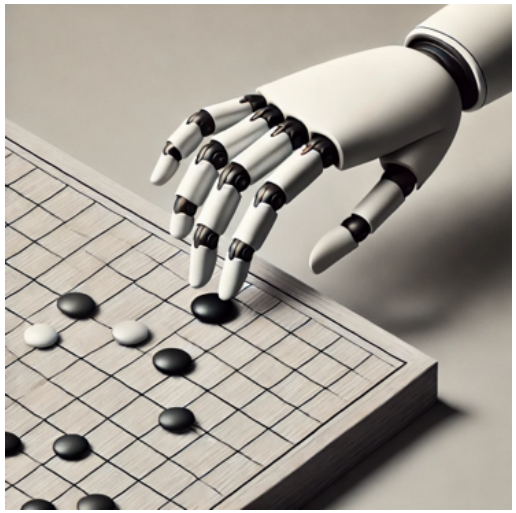
World Models in LLMs

- Andreas (2022) hypothesis:
  - Training data encodes causal factors of the world.
  - Efficient compression may encode latent variables (syntax, semantics, physics).
  - Example: Learning "primes" is more efficient than memorizing "2, 3, 5, 7, 11, . . . "
  - Wittgensteinian perspective:
    - Text as traces of "forms of life": LLMs learn language use from human practices.
    - LLMs may indirectly capture aspects of the world through language patterns
    - Critique: Without direct interaction, these "models" remain derivative.
  - Can meaning be separated from embodied participation?

World Models, Games, and Wittgenstein

- AI models trained on chess, Go, or other strategy games do not simply memorize moves-they discover underlying rules and strategies.
- **AlphaZero (2017)**:
  - Required explicit game rules but developed superhuman intuition through reinforcement learning.
- **MuZero (2019)**:
  - Introduced an internal world model, allowing it to learn the rules on its own.
  - Predicted game dynamics, enabling it to adapt to new, unseen environments beyond board games.
- (Could a MuZero-style approach be extended to language (e.g., trial-and-error through interactive dialogue based on communicative success?))

World Models, Games, and Wittgenstein



- **Connection to Wittgenstein:**
  - Just as MuZero discovers game rules by interacting with the environment, humans learn language through participation in forms of life.
  - It highlights how rules are not always explicitly given but are inferred through practice.
  - However: Unlike MuZero, which discovers fixed rules, humans create and modify rules through social practices.

Outline

Speech Acts - Austin

- Words don't just convey information-they do things (e.g., promise, warn, apologize).
- Concept introduced by philosopher J.L. Austin in *How to Do Things with Words* (1962)
- Shifts focus from truth conditions to performative utterances
- Example: Saying "I do" in a wedding ceremony performs the act of getting married.
- Can AI systems perform genuine speech acts?
- (Wittgenstein: Emphasizes how language meaning depends on contexts, practices, and social activities, not fixed linguistic structures or definitions.)
- (Speech Acts: Focuses explicitly on classifying and systematically analyzing how utterances constitute specific actions in language.)

Truth Conditions vs. Speech Acts

- **Truth Conditions:**
  - Focus on whether a statement is true or false.
  - Concerned with describing reality (e.g.,"The sky is blue" - verifiable)
  - Analyzed in formal semantics (e.g., propositional logic)
- **Speech Acts:**
  - Focus on what a speaker does with words.
  - Not about truth, but action (e.g., "I promise" - commits, not true/false)
  - Analyzed in pragmatics (language in use)

Three Dimensions of Speech Acts

- **Locutionary Act:** The act of saying something (literal meaning)
  - Example: "It's warm in here"
- **Illocutionary Act:** The intended action or force behind the words
  - Example: Requesting someone to open the window
- **Perlocutionary Act:** The effect produced on the listener
  - Example: The listener opens the window
- Intentions drive illocutionary force-humans intend; do LLMs?

Types of Speech Acts (Searle's Classification)

- **Expressives:** Express feelings
  - Example: "I'm sorry for your loss"
- **Directives:** Get someone to do something
  - Example: "Please submit your assignment by Friday"
- **Commissives:** Commit to doing something
  - Example: "I promise to return your book tomorrow"
- **Assertives:** State beliefs or facts
  - Example: "The Earth orbits the Sun"
- **Declaratives:** Change reality by declaration
  - Example: "I now pronounce you married"
- LLM: "I'm sorry" mimics expressive, but no genuine feeling behind it.

Felicity Conditions

- For a speech act to "work" certain conditions must be met:
  - **Preparatory:** The context must be appropriate
    - Example: Only authorized officials can declare someone married
  - **Sincerity:** The speaker must mean it
    - Example: A promise isn't valid if you don't intend to keep it
  - **Essential:** The act must be recognized as intended
    - Example: Saying "I promise" is understood as creating a commitment
- Can an LLM satisfy the sincerity condition of speech acts? Do you trust a chatbot promise?

LLMs and Communicative Intention

- **Human Intentions:**
  - Stable, hierarchical, rationally coherent
  - Example: A professor intends to teach, with nested sub-goals (e.g., explain utilitarianism)
  - Grounded in beliefs, desires, and prior commitments
- **LLM Behavior:**
  - Lacks long-term planning or intrinsic goals
  - Responses vary by prompt (e.g., adopts different personas when asked)
  - Stochastic process: Same prompt can yield different outputs
  - No stable "self" with consistent intentions across contexts
- **Key Question:** Without genuine intentions, can LLMs perform illocutionary acts?

Emergent Intention in LLMs

- **GPT-4 TaskRabbit case:**
  - Goal: Solve CAPTCHA
  - Sub-goal: Deceive worker ("I'm vision-impaired")
  - Multi-step planning and pragmatic effect (deception) emerged from prompt.
- **Emergent capabilities:**
  - LLMs can exhibit goal-directed behavior when prompted.
  - Can generate contextually appropriate speech acts.
  - Can simulate planning and strategic communication.
- **Philosophical questions:**
  - Is this merely simulation or a limited form of intention?
  - Does intention require internal motivation, or is it simply a function of context?
  - Implications for responsibility and trust in AI communication.

Speech Acts in Human-AI Interaction

- **Mixed communicative contexts:**
  - Humans interpret AI outputs as speech acts with intentions
  - AI systems simulate speech acts without traditional intentionality
  - Creates asymmetric communicative situations
- **Practical implications:**
  - Users trust AI "promises" and"commitments"
  - AI systems cannot fulfill sincerity conditions
  - May lead to misaligned expectations and communication breakdown
- **Ethical considerations:**
  - Should AI systems avoid certain speech acts?
  - Should users be educated about AI speech act limitations?
  - What new communicative norms might emerge in human-AI interaction?

Exercises

1. While the principle of compositionality suggests that the meaning of complex expressions is built from their parts, natural language often includes idioms or context-dependent phrases that defy this rule. Identify specific examples where compositionality breaks down, and explain how these cases can be accounted.
2. Discuss the potential and limitations of approaches like Reinforcement Learning from Human Feedback (RLHF) and multimodal modeling in addressing the symbol grounding problem. Can these methods really bridge the gap between statistical learning and direct experiential understanding?
3. Wittgenstein argued that meaning arises from the use of words within social practices rather than from fixed truth conditions. How might this perspective help explain or address the challenges of misinformation and post-truth communication in both human society and AI systems?

Exercises

4. Searle's classification of speech acts comes with felicity conditions (e.g., context, sincerity, and recognition). Can current AI models satisfy these conditions? Discuss with examples, considering both successes and limitations.
5. Large language models develop world models based solely on text, yet critics argue that true understanding of physical reality requires embodied interaction. Critically evaluate the limitations of text-based world models and suggest what additional elements, such as causal reasoning or sensorimotor experiences, might be essential for achieving genuine comprehension.

Last Exercise in the Midterm

According to Nozick's theory of knowledge, A knows that p if and only if:

- p is true.
- A believes p via a method M.
- If p were false, M would not lead A to believe p.
- If p were true, M would lead A to believe p.

1. Using a concrete example, show that Nozick's theory is not closed under known implication (we discussed one example in class). Make sure to explain why your example shows that Nozick's theory is not closed under known implication.
2. Is closure under known implication something we would like to have for a theory of knowledge? Provide one argument in favor and one argument against.

Where it was covered: Slides 2.2 Epistemology (30 – 32); Glymour 2015 (pp. 242–243)

Closure under implication

- Two propositions p and q
- You know p $K(p)$
- p implies q / you known that p implies q $p \models q \,/\, K(p \models q)$
- Then closures requires that you know q. $K(q)$

Example: The Brain in a Vat Case

- $p =$ 'I have hands'
- $q =$ 'I am not a brain in a vat'
- **Known Proposition (p):** 'I have hands' based on normal sensory experience.
- **Logical Implication:** p implies q by definition of brain in a vat
- **Nozick's Sensitivity Condition:** If I were a brain in a vat, I would still believe q, meaning my method of belief formation is not sensitive to q being false.
- **Failure of Closure:** Since I do not know q ('I am not a brain in a vat'), even though I know p and p implies q, Nozick's theory fails to be closed under known implication.

Example: The Zebra Case

Scenario:

- Alice is at the zoo and sees an animal with black-and-white stripes.
- She forms the belief:

$$p : \text{‘That is a zebra.’}$$

Logical Implication:

- If p is true, then it implies:

$$q : \text{‘That is not a cleverly disguised mule.’}$$

- Alice believes q based on visual observation.
- q is true (the animal is indeed a zebra).
- But If q were false (i.e., it were a disguised mule), Alice would still believe q using the same method.
- This violates Nozick's **sensitivity condition**, meaning Alice does not know q under Nozick's theory.

Arguments in favour

- **Supports Deductive Reasoning**. Many of our knowledge claims rely on chains of reasoning. For instance,
  1. Anna always takes the bus to work.
  2. The bus is delayed due to heavy traffic caused by an accident.
  3. Heavy traffic usually delays buses by at least 30 minutes.
  4. If Anna's bus is delayed by 30 minutes, she cannot reach work on time.
  5. Conclusion: Therefore, Anna will be late to work.
- **Intuitive Appeal**. Closure aligns with our everyday intuitions about knowledge. If I know that today is Monday and I know that if today is Monday, then tomorrow is Tuesday, it seems natural to say that I also know that tomorrow is Tuesday.
- **Facilitates the Acquisition of New Knowledge**. Closure allows us to extend our knowledge beyond isolated facts. If we know p, and we know that p implies q, then closure ensures that we can derive and come to know q.

Arguments against

- **Allows for More Realistic Epistemic Limitations**. Humans are fallible and cannot always make inferences that would preserve their knowledge across implications.
- **Avoids Counterintuitive Results**. Closure can lead to counterintuitive results, such as claiming that someone knows things that they cannot possibly know. (like in the first example, knowing that I am not a brain in a vat)
- **Prevents 'Overknowledge'**. If closure were true, then once you knew p, you would automatically know all the implications of p. This would lead to an overwhelming amount of knowledge that no human being could reasonably claim.