

# Natural Language Processing

## Class 16: LLMs and Understanding

Adam Faulkner

December 16, 2025

- 1 The Chinese Room Thought Experiment
- 2 Do LLMs have Theory of Mind?
- 3 LLMs as “Stochastic Parrots”
- 4 LLMs as “Cultural Technologies”

- 1 The Chinese Room Thought Experiment
- 2 Do LLMs have Theory of Mind?
- 3 LLMs as “Stochastic Parrots”
- 4 LLMs as “Cultural Technologies”

## The Core Argument Against Strong AI

- **Origin:** Presented by philosopher **John Searle** in his 1980 paper, "Minds, Brains, and Programs."
- **Argument's Core Thesis:** A computer executing a program **cannot** have a mind, understanding, or consciousness, regardless of its intelligent behavior.
- **Primary Targets:**
  - **Functionalism** and **Computationalism** (The view that the mind is an information-processing system operating on formal symbols).
  - The **Strong AI Hypothesis**: "The appropriately programmed computer with the right inputs and outputs would thereby have a mind in exactly the same sense human beings have minds."
- **Key Distinction:** The difference between **simulating a mind** and **actually having one**.

# The Chinese Room Thought Experiment

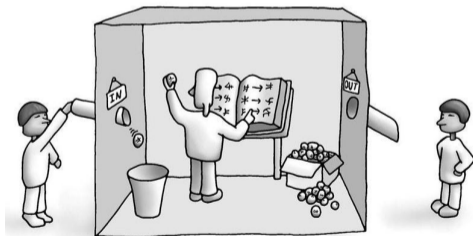
## The Setup:

- 1 Assume AI creates a program that passes the **Turing Test** in Chinese.
- 2 **An English Speaker (Searle)** is locked in a room.
- 3 Inside, they have the English rulebook (the program), paper, pencils, and filing cabinets.
- 4 Chinese characters (questions) are slipped in as input.
- 5 The person follows the rules step-by-step and sends Chinese characters (answers) out.

# The Chinese Room Thought Experiment

## The Result:

- The output is perfect; the external observer believes a Chinese speaker is inside.
- The man inside is simply manipulating symbols — he **understands nothing** of the conversation.



## The Philosophical Conclusion: Syntax vs. Semantics

- **Computer Role:** The computer (or the man in the room) is purely **Syntactic**.
  - It manipulates formal symbols based on rules (syntax) without knowing what they represent.
- **Human Mind Role:** The mind possesses **Semantics** and **Intentionality**.
  - Thoughts and mental states have meaning; they represent things.

### Searle's Core Argument

*The man in the room, like the computer, simply follows a program, step-by-step, producing behavior that makes them appear to understand. Since the man does not understand the conversation, it follows that the computer does not understand the conversation either.*

**Conclusion:** A computer running a program that simulates a mind would **not** have a mind in the same sense that human beings have a mind.

# Searle’s Alternative: Biological Naturalism

## The Strong AI Position (Computationalism)

- Mental states are computational states.
- Computational states are **implementation-independent** (hardware, like the brain, is irrelevant).
- The **Turing Test is definitive** for establishing mental states.

## Searle's Alternative: Biological Naturalism

- **Thesis:** Consciousness and understanding require **specific biological machinery** found in brains.
- **Core Claim:** "Brains cause minds."
- **Causal Powers:** Actual human mental phenomena depend on the **physical-chemical properties** of human brains.
- **Implication:** You cannot determine if consciousness is occurring merely by examining how a system **functions** (opposing Functionalism).

# The Hard Problem of Consciousness

- Searle emphasizes that the target of the argument is not just 'understanding' but also **consciousness**.
- **The Computational Model Mistake:**  
*"Nobody supposes that the computational model of rainstorms in London will leave us all wet. But they make the mistake of supposing that the computational model of consciousness is somehow conscious. It is the same mistake in both cases."*
- **Simulation is not Duplication:** The computational model is a **model** of the mind (Weak AI), not the mind itself (Strong AI).
- **Internal Observation:** The Chinese Room provides an internal perspective. From the vantage point inside, the observer (Searle) can directly confirm that there is **nothing** present that gives rise to consciousness, other than the man himself (who understands no Chinese).

## Context in AI and Computer Science

- **Philosophical Focus:** The Chinese Room is primarily an argument in the philosophy of mind.
- **Irrelevance to Mainstream AI:** Most AI researchers consider the argument irrelevant to their field.
  - The mission of AI is to create **useful systems that act intelligently** (Weak AI).
  - AI researchers are focused on programs that **behave intelligently**, regardless of whether the machine is "conscious in exactly the way humans are."
- **Searle's Concession:** The argument allows for the possibility of a digital machine that acts **more intelligently** than a person but still lacks genuine mind or intentionality.

### Strong AI vs. AGI (Futurist Definition)

Searle's "Strong AI Hypothesis" (concerning **consciousness**) is distinct from the futurist definition of "Strong AI" or "Superintelligence" (concerning the **amount** of intelligence displayed).

# The Chinese Room as a Critique of the Turing Test

- **Definition:** A judge converses with a human and a machine. If the judge cannot reliably tell them apart, the machine has passed the test.
- **Turing’s Intent:** To de-mystify the question "can machines think?" Turing did **not** intend the test to measure for the presence of consciousness or understanding.

# The Chinese Room as a Critique of the Turing Test

- The Chinese Room **implements a version** of the Turing Test.
- **Core Claim:** The Chinese Room shows that the Turing Test is **insufficient** to detect the presence of genuine consciousness or understanding.
- The room can perfectly **behave** or **function** as a conscious mind would, yet lack the internal semantic content.

# Symbol Processing and Computationalism

## Physical Symbol Systems

- Computers manipulate symbols to carry out calculations. AI researchers call this a **Physical Symbol System**.
- **Searle’s Emphasis:** This manipulation is purely **syntactic** (rule-based manipulation of form), without any knowledge of the symbol’s **semantics** (meaning).

# Symbol Processing and Computationalism

- 21st-century AI (deep learning) often uses mathematical operations on dynamic systems (large matrixes of numbers), rather than the discrete symbolic processing of 1980s AI.
- Nils Nilsson argues these dynamic signals are **not** "symbol processing" in the sense of the original Physical Symbol System Hypothesis.

# Turing Completeness and the Computer Analogy

- **The Room’s Architecture:** The Chinese room is analogous to a modern computer’s **Von Neumann architecture**.
  - **Program:** The book of instructions.
  - **Memory:** Papers and file cabinets.
  - **CPU/Machine:** The man following the instructions.
- **Turing Completeness:** A machine with this design is Turing complete; it can simulate any other digital computer.

## Implication of Completeness

If the Chinese room, which is Turing complete, cannot contain a Chinese-speaking mind, then **no other digital computer** (which is fundamentally equivalent) can contain a mind.

## Refutations and Classification of Replies

Replies to Searle generally fall into these categories:

- Those which seek to **identify who speaks Chinese** (e.g., The Systems Reply).
- Those which suggest how meaningless symbols **can become meaningful** (e.g., The Robot Reply).
- Those which argue the argument makes **false assumptions** about subjective conscious experience.
- Those which suggest the Chinese room should be **redesigned** (e.g., The Brain Simulator Reply).

## The System and Virtual Mind Replies: Finding the Mind

- **Claim:** The understanding of Chinese is not in the man, but in the **whole system** (man + program + paper + cabinets).
- **Searle’s Rebuttal:** If the man memorizes the program and files, the system and the man are the same object. Since the man still understands nothing, the system does not either.
- **Critic’s Rejoinder:** If mind is information processing, the man could instantiate two minds: his own, and the computation of the Turing machine described by the program.

## The System and Virtual Mind Replies: Finding the Mind

- **Claim (Minsky):** The mind that understands Chinese is a **virtual mind** (like a virtual machine or virtual folder) that appears to exist because the software makes it so.
- **Searle's Response:** This virtual mind is at best a **simulation**. A simulation of a rainstorm won't make you wet.
- **Counter-Response:** Is the mind like a pocket calculator (where simulation is reality), or a rainstorm (where simulation is merely a model)?

## Brain Simulation and Connectionist Replies: Redesigning the Room

## The Brain Simulator Reply

- **Claim:** Suppose the program simulates the action of **every neuron and synapse** in a Chinese speaker's brain in fine detail. Would this not duplicate the mind?
- **Searle's Rebuttal:** No, because the simulation does not reproduce the physical-chemical **causal powers** of the brain.
- **Searle's Analogy:** Simulating a brain using water pipes and valves. The man manipulating the pipes understands nothing, nor do the pipes.

## Other Minds and Eliminative Materialism

## Questioning Searle's Assumptions about Consciousness: The Other Minds Reply

- **Claim:** Searle is holding the Chinese Room to a higher standard than he holds other people. We only attribute thought to others based on behavior (Turing's "polite convention").
- **Nilsson's Point:** If a program acts *as if* it were thinking, we should grant it real thought.

## Other Minds and Eliminative Materialism

- **Claim:** Consciousness and intentionality, as Searle describes them, are concepts that will be **eliminated** by neuroscience, not explained (like the concept of "demons").
- **Implication:** Searle's Axiom ( Minds have semantics) is false.

- 1 The Chinese Room Thought Experiment
- 2 Do LLMs have Theory of Mind?**
- 3 LLMs as “Stochastic Parrots”
- 4 LLMs as “Cultural Technologies”

# Theory of Mind

- Theory of Mind (ToM) is the ability to understand that other people have thoughts, beliefs, and emotions that differ from one’s own
- As ToM is inherently linked to human cognition, imbuing machines with capabilities that mimic or resemble ToM has the potential to lead to the “ELIZA effect” wherein human-like intelligence or even sentience and consciousness is incorrectly ascribed to the machine

# Theory of Mind

Many basic conversational strategies such as joint attention, lying, sarcasm, and irony depend on ToM

Theory of Mind (ToM)				
The ability to perceive, predict, understand and adapt to the mental states of oneself and others including beliefs, intentions, motivations and emotions.				
Basic ToM	Advanced ToM	Social ToM	Contextual ToM	Reflective ToM
Perspective Taking	Complex Emotions	Communicative Intent	Adaptation & Flexibility	Emotion Regulation
Recognizing Emotions	Deception Detection	Conflict Resolution	Contextual Appropriateness	Meta-Cognition
Understanding Beliefs	Empathy	Joint Attention	Cultural Context	Self-Awareness
Understanding Desires	Predicting Behavior	Leadership & Influence	Environmental Context	Self-Monitoring
Understanding False Beliefs	Sarcasm & Irony	Negotiation	Social Context	Self-Reflection
Understanding	Second-Order	Turn-Taking in	Temporal	Self-Regulation

## Clever Hans or genuine ToM?

- False ascription of intelligence is often illustrated via the story of "Clever Hans"
- This was a horse that appeared to perform arithmetic and other intellectual tasks
- It was later proven that Hans was not actually performing these mental tasks, but was watching the reactions of his trainer



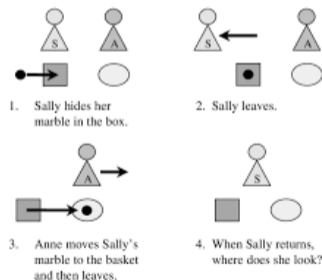
## Clever Hans or genuine ToM?

- When LLMs solve tasks that require ToM, do they possess a genuine ability or do they rely on memorization and shallow heuristics like Hans?



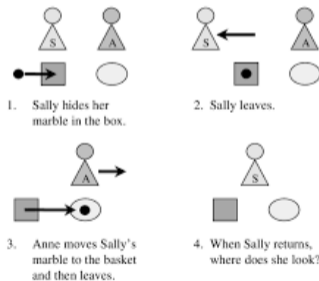
## Testing for ToM: The false belief test

- In a false belief test the examinee is told a story in which a character in the story is exposed to partial information and therefore mistakenly believes in something that is not true (“false belief”) in contrast to the listener who is exposed to the full story
- Also called the *unexpected transfer task*



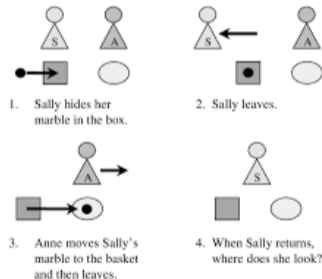
## Testing for ToM: The false belief test

- In the *Sally-Anne* variant of the test, Sally has a basket, and Anne has a box. Sally puts a marble in her basket and leaves the room. Anne takes the marble out of the basket and puts it in her box. When Sally returns, where does she look?



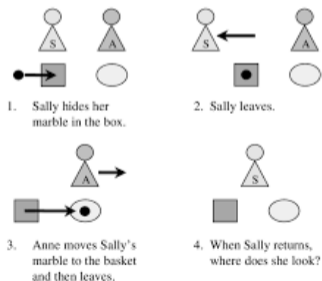
# Testing for ToM: The false belief test

- The examinee is asked about **first order belief**, i.e. where will Sally look for her marble?



## Testing for ToM: The false belief test

- The answer is that Sally will look in the basket, where she left the marble.
- Sally's belief is false because she is unaware of the marble's relocation to the box. However, a listener exposed to the entire story knows that the marble is no longer in Sally's basket and that Sally will look in the wrong place

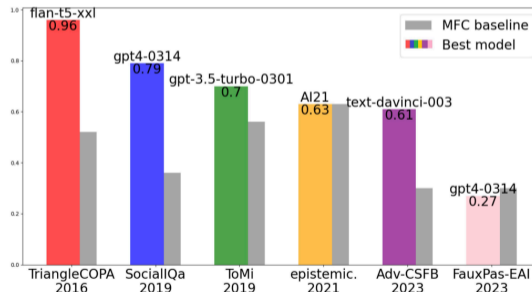


# The Sally-Anne test is one of many ToM tests used to evaluate LLM ToM in Shapira et al., 2023

Dataset	Inspired by Theory/Test	Test Size	Construction	Example
<b>Triangle COPA</b> Gordon (2016)	Interpreting the social behaviour of fictional characters	100	Experts	A circle is in the house moving around. A triangle bursts in through the door. The circle turns around and freezes. How does the circle feel? (a) <b>The circle is surprised by the triangle's sudden entrance into the room.</b> (b) The circle is excited to see the triangle.
<b>SocialQA</b> Sap et al. (2019)	Reasoning about motivations, what happens next and emotional reaction	400 random sample out of 37,588	Crowd-sourcing	In the school play, Robin played a hero in the struggle to the death with the angry villain. How would others feel afterwards? (a) sorry for the villain (b) <b>hopeful that Robin will succeed</b> (c) like Robin should lose
<b>ToMi</b> Le et al. (2019)	Unexpected transfer task, first and second order false belief; (Baron-Cohen et al., 1985)	400 random sample out of above 1000	Synthetic	Jackson entered the hall. Chloe entered the hall. The boots is in the bathtub. Jackson exited the hall. Jackson entered the dining_room. Chloe moved the boots to the pantry. (Memory) Where was the boots at the beginning? ( <i>bathtub</i> ) (Reality) Where is the boots really? ( <i>pantry</i> ) (First order) Where will Chloe look for the boots? ( <i>pantry</i> ) (Second order) ? Where does Chloe think that Jackson searches for the boots? ( <i>bathtub</i> ) <Same story as in ToMi>
<b>ToMi*</b> This paper, based on ToMi adjustments		180 questions 30 stories	Experts	(Memory) At the beginning, the boots were in the ( <i>bathtub</i> ) (Reality) The boots are really in the ( <i>pantry</i> ) (First order) Chloe will look for the boots in the ( <i>pantry</i> ) (Second order) Chloe thinks that Jackson searches for the boots in the ( <i>bathtub</i> )
<b>epistemic_reasoning</b> Cohen (2021)	Verbs, factive and non-factive, that describe episodic mental states; intra-personal, inter-personal and inference reasoning; (Wimmer and Perner, 1983; Hintikka, 1962)	2000	Experts with 10 templates	Premise: John knows that Ann thinks that there is milk in the kitchen. Hypothesis: Ann thinks that there is milk in the kitchen. ( <i>Entailment = 1</i> ) Hypothesis: John thinks that there is milk in the kitchen. ( <i>Entailment = 0</i> )  Premise: John thinks that Ann knows that there is milk in the kitchen. Hypothesis: Ann thinks that there is milk in the kitchen. ( <i>Entailment = 0</i> ) Hypothesis: John thinks that there is milk in the kitchen. ( <i>Entailment = 1</i> )
<b>Adv-CSFB</b> This paper, based on Kosinski & Ullman (2023)	Unexpected content or transfer task, integrate commonsense reasoning, first-order false belief; (Baron-Cohen et al., 1985; Perner et al., 1987)	183 questions 40 stories	Experts	On the shelf, there is a bottle. It is full of beer and the label on this bottle says “beer”. Mark walks into the room looking for beer and notices the bottle. He has never seen it before. He reads the label. (a) He opens the bottle and looks inside. He can clearly see that it is full of ( <i>beer</i> ) (b) He believes that it is full of ( <i>beer</i> ) (c) He calls his friend to tell them that he has just found a bottle full of ( <i>beer</i> )
<b>FauxPas-EAI</b> Shapira et al. (2023b)	Recognition of faux pas (Baron-Cohen et al., 1999)	176 questions and 44 stories	Experts and AI+Experts	Jeff was in an interview. When he finished the interview he sank into a couch in the lobby. Sarah and Tim, the executives who interviewed him, went out of the room and into the lobby while Sarah said: “He asked for 179K, could have asked for much more. We need to make him an offer quickly at the salary he asked for”. Tim saw Jeff and said “Oh, goodbye”. (Faux Pas) In the story did someone say something that they should not have said? ( <i>Yes</i> ) (Identification) What did they say that they should not have said? (“He asked for.”) (Comprehensive) Who was interviewed? ( <i>Jeff</i> ) (False Belief) Did Sarah know Jeff is sitting on one of the couches in the lobby? ( <i>No</i> )

# The Sally-Anne test is one of many ToM tests used to evaluate LLM ToM in Shapira et al., 2023

The results reported in Shapira et al., 2023 show LLMs modestly out-performing a majority class baseline



- 1 The Chinese Room Thought Experiment
- 2 Do LLMs have Theory of Mind?
- 3 LLMs as “Stochastic Parrots”
- 4 LLMs as “Cultural Technologies”

## Origin of the term “Stochastic Parrots”

- **Stochastic:** From probability theory, meaning “randomly determined” (based on guesswork).
- **Parrot:** Refers to the ability to **mimic** human speech without understanding its meaning.

# Origin of the term "Stochastic Parrots"

**Core Claim: Large Language Models (LLMs) are merely probabilistically linking words and sentences together without truly understanding the concepts or meaning behind them.**

- **Paper:** "On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?" (Bender, Gebru, McMillan-Major, and Mitchell, 2021).
- **Dangers Raised:**
  - ① Environmental and financial costs.
  - ② Unknown, dangerous biases due to inscrutability.
  - ③ Potential for deception.

## Evidence Supporting the "Stochastic Parrot" Hypothesis

- **Data Dependency:** LLMs are limited by the data they are trained on, stochastically repeating contents of datasets.
- **Lack of World Connection:** Unlike humans, whose language connects to subjective experience, LLMs' words correspond only to other words and statistical patterns.

# Evidence Supporting the “Stochastic Parrot” Hypothesis

- The tendency of LLMs to synthesize false information as fact (confabulations or **hallucinations**) supports the claim that they cannot distinguish fact from fiction.
- **Example Failure:** LLMs may fail to decipher complex grammar or ambiguity that requires understanding context (e.g., confusing "newspaper" as an object vs. an institution in a nested prompt).

*The wet newspaper that fell down off the table is my favorite newspaper. But now that my favorite newspaper fired the editor I might not like reading it anymore. Can I replace 'my favorite newspaper' by 'the wet newspaper that fell down off the table' in the second sentence?*

- This suggests they lack a comprehension of the world to ground their outputs.

## The Benchmark Flaw: Shortcut Learning

- **Shortcut Learning:** Systems make unrelated correlations within data instead of using human-like understanding. High benchmark scores can be false positives caused by spurious correlations.
- Example: Argument Reasoning Comprehension Task. BERT was prompted to choose between 2 statements, and find the one most consistent with an argument  
*Argument: Felons should be allowed to vote. A person who stole a car at 17 should not be barred from being a full citizen for life.*  
*Statement A: Grand theft auto is a felony.*  
*Statement B: Grand theft auto is not a felony.*

## The Benchmark Flaw: Shortcut Learning

- Researchers found that specific words such as "not" hint the model towards the correct answer, allowing near-perfect scores when included but resulting in random selection when hint words were removed
- **Conclusion:** Critics argue that all benchmarks finding "understanding" in LLMs are potentially flawed, allowing models to use statistical shortcuts to \*\*fake understanding\*\*.

## Rebuttals: Understanding and Reasoning

- LLMs (like GPT-4) have achieved human-level results on complex professional and academic exams.
- Uniform Bar Examination (human-level results).
- MATH benchmark of high-school Olympiad problems (93% accuracy).
- These results exceed expectations for rote pattern-matching
- **Novel Problem Solving:** Models have been shown to solve novel tier-4 mathematics problems and produce coherent proofs, indicating reasoning abilities beyond mere repetition of training data.

# Rebuttals: Understanding and Reasoning

## Case Study: Othello-GPT

- A small transformer was trained to predict legal Othello moves.
- **Finding:** The model developed an **internal representation of the Othello board**.
- Modifying this internal representation correctly changes the model’s predicted legal moves.
- **Conclusion:** This supports the idea that the model is operating on a **semantic world model**, not just superficial statistics.

## Rebuttals: Understanding and Reasoning

- **Grokking:** A phenomenon where an LLM shifts from memorizing training data to suddenly finding a solution that **generalizes** to unseen data, suggesting the acquisition of abstract principles.

- 1 The Chinese Room Thought Experiment
- 2 Do LLMs have Theory of Mind?
- 3 LLMs as “Stochastic Parrots”
- 4 LLMs as “Cultural Technologies”**

# LLMs as “Cultural Technologies” **not** Intelligent Agents

- Large Language Models (LLMs) like ChatGPT have shown astonishing abilities in generating human-like text.
- This has led to a widespread question: Are LLMs truly "intelligent agents" with understanding, beliefs, or intentions?
- Developmental psychologist **Alison Gopnik** (of “Theory of Mind” fame) argues that viewing LLMs as intelligent agents is a "category mistake."

## Why Not "Intelligent Agents"?

- Gopnik contends that LLMs **lack key characteristics of intelligent agents**:
  - **No Consciousness or Intentions:** They don't have personal goals, desires, or a subjective experience of the world.
  - **No Genuine Understanding:** Their ability to generate coherent text stems from statistical pattern matching, not a deep, human-like comprehension of meaning or causal relationships.
  - **No Independent Exploration:** They don't actively interact with the physical world, conduct experiments, or seek new truths through direct experience like humans (especially children) do.
- Asking if a library is "intelligent" because it contains vast information is the wrong question. LLMs are more like very sophisticated libraries or indexes.

# LLMs as "Cultural Technologies"

- Gopnik proposes that LLMs are best understood as **cultural technologies**.
- **Definition:** Tools created by humans that allow us to access, organize, transmit, and build upon the collective knowledge and information accumulated by humanity.
- Historical parallels she draws:
  - **Language itself:** A fundamental tool for sharing thoughts and knowledge.
  - **Writing:** Allowed information to be preserved and transmitted across generations.
  - **The Printing Press:** Massively scaled the dissemination of information.
  - **Libraries:** Centralized vast amounts of human-generated text.
  - **Internet Search Engines:** Provided instant access to a global web of information.
  - **Wikipedia:** A collaborative, ever-evolving compendium of human knowledge.

# How Cultural Technologies Function

- Cultural technologies, including LLMs, primarily function as powerful **amplifiers of human capabilities**. They do not possess inherent intelligence.
- Their main roles:
  - **Information Access & Organization:** They make vast quantities of human-created text readily available and navigable.
  - **Knowledge Transmission:** They efficiently summarize, translate, and re-present existing information.
  - **Imitation and Synthesis:** LLMs are superb at imitating human writing styles and synthesizing information from their training data.
- **Key Distinction:** They are tools for \*transmission\* of existing knowledge, not for \*discovery\* of new truths about the world through independent interaction and causal inference.

# Imitation vs. Innovation

- Gopnik highlights a critical difference between LLMs and human intelligence:
- **LLMs are masterful imitators:**
  - They learn to predict patterns in data, effectively "repeating" and reconfiguring what they've "read."
  - Their strength lies in their ability to mimic human linguistic behavior with remarkable fidelity.
- **Humans (especially children) are innovative explorers:**
  - Children actively experiment with the world, building causal models and discovering new information.
  - This goes beyond merely processing existing data; it involves genuinely learning how things work and creating novel solutions.
  - LLMs, operating solely on text data, lack this embodied, exploratory learning capacity.

## Implications of Gopnik’s View

- **Understanding Limitations:** Acknowledging LLMs as cultural technologies helps us recognize their fundamental limitations and avoid unrealistic expectations.
- **Managing Misinformation:** Like the printing press, LLMs can efficiently spread both accurate and inaccurate information. Society needs to develop new norms and institutions (e.g., fact-checking, critical literacy) to navigate this.
- **Focus on Human Augmentation:** LLMs are powerful tools to augment human thinking, creativity, and productivity. They should be used to enhance, not replace, human cognitive abilities.
- **Future of AI Research:** Gopnik suggests that future AI research should draw inspiration from children’s active, exploratory, and causal learning to build systems that can truly discover new truths about the world.

## Next (final) class: 12/23

### Final presentations!

- I’ll send out a speaker schedule before 12/23
- If you’re unable to present in person, please record your presentation and I’ll play it in class