Language of Thought for a Probabilistic Theory of Mind

Aidan J. Wagner
April 21, 2026

Computational theories of mind, as the name suggests, treat the mind as a computer; a computer that manipulates thoughts in an algorithmic fashion. The Language of Thought (LoT) hypothesis is one of these computational theories that has been taken as an account of how thought is structured. The LoT elegantly explains how an infinite number of distinct thoughts can be generated from a finite set of primitives by combining these primitives into sentences in a linguistic way. But any theory of the mind in general must capture the mind's epistemic life under uncertainty: how agents represent graded confidence and rationally update their attitudes in light of evidence. The digital architecture of the classical LoT appears to be at odds with the probabilistic constraints demanded by a Bayesian approach to epistemology. The classical LoT hypothesis and its syntactic manipulation of discrete symbols lack a certain "richness" necessary to encode the continuous degrees of partial belief demanded by probabilism. Therefore, the cognitive architecture of the LoT hypothesis requires a radical update to accurately model how the mind represents and computes these degrees of belief.

Bayesian Epistemology & Probabilism

Epistemology is focused on *representational attitudes*, which are "ways that agents take the world to be" [3, p. 3]. In other words, they are attitudes that assert descriptive claims about the veracity of propositions. (This is just a more general class of mental states that includes belief.) Bayesian epistemology narrows this further to what are called *credences*, or degrees of partial belief/degrees of confidence. These credences are assigned to propositions (truth-evaluable statements). The full set of numerical credences an agent assigns to the propositions in their language is known as a *credence distribution*. *Probabilism* is the philosophical thesis that rationality requires an agent's credence distribution to satisfy Kolmogorov's three axioms, which define mathematical probability. A distribution that satisfies the three axioms is called a *probability distribution*. Kolmogorov's three axioms are as follows:

1. **Non-Negativity:** For any proposition P, the assigned credence must be greater than or equal to zero ($cr(P) \geq 0$).

2. **Normativity:** For any logical tautology T, the assigned credence must be exactly equal to one ($cr(T) = 1$).

3. **Finite Additivity** [1]: For any mutually exclusive propositions P and Q, the credence assigned to their disjunction $(P \vee Q)$ must be exactly equal to the sum of their individual credences $(cr(P \vee Q) = cr(P) + cr(Q))$.

In practice, adopting probabilism means modeling an agent's doxastic mental states on a continuous graded scale rather than discrete categorical assignments. The axioms are normative rules for how one *should* manage their uncertainty, not descriptive requirements for possessing uncertainty in the first place. By constraining rational belief in this way, a good portion of the probability calculus becomes available to model the reasoning and behavior of rational agents. One such tool from the foundations of probability theory that has proven particularly useful is *Bayes' theorem*, which, in the context of credence distributions, provides the ability to model how credences change in the face of new information via probabilistic conditioning:

The Rule of Conditionalization: When an agent learns a new piece of information E between time t_i and a later time t_j, conditionalization demands that their unconditional credence in any hypothesis H at time t_j must be exactly equal to the conditional credence held at t_i for H given E. Formally, this is denoted as $cr_j(H) = cr_i(H|E)$ [3, p. 92].

To actually compute this conditional probability, Bayes' theorem is applied, which yields:

$$cr(H|E) = \frac{cr(E|H) \cdot cr(H)}{cr(E)}$$

This models the updating of beliefs by decomposing the resulting credence into three pieces:

- **The Prior** $cr(H)$**:** the agent's unconditional degree of belief in the hypothesis before the new evidence is collected.
- **The Likelihood** $cr(E|H)$**:** the probability of observing the specific evidence E, assuming that the hypothesis H is true.
- **The Expectedness of the Evidence** $cr(E)$**:** the agent's prior unconditional credence that the evidence E would be observed at all, regardless of the truth-value of the hypothesis.

The Language of Thought Hypothesis

The language of thought hypothesis proposes that conceptual thinking occurs in an internal, language-like medium [2, p. 280]. This mental language contains "words" that combine to make

[1] Note: in most mathematical contexts, the third axiom is strengthened to Countable Additivity, which is necessary to prove a lot of the asymptotic behavior that makes up much of modern probability theory. However, we do not require countable additivity, and some Bayesian philosophers (Rush Stewart) reject the idea that credences need to be countably additive. So, here I defer to Occam's razor and use the weaker of the two claims.

"sentences" according to grammatical rules. The LoT hypothesis is a computational theory of mind, positing that reasoning is fundamentally a process of symbol manipulation governed by algorithms [2, p. 281].

The classical LoT hypothesis rests on three core claims:

Cognitive processes consist of causal sequences of tokenings of internal representations. Mental states like beliefs and desires involve literal physical instances (tokens) of mental representations in the brain. Thinking is simply the causal sequencing of these representations [2, p. 281].

Mental representations have combinatorial syntax and semantics. Complex mental representations are built out of atomic symbols combined according to linguistic rules. The semantics of a complex thought is a function of the meanings of its atomic symbols and the structure in which they are arranged [2, p. 281].

Mental operations are causally sensitive solely to the syntactic structure of the symbol. The mind manipulates these symbols based on their formal structural properties (their "shape" or syntax) rather than their semantic meaning. The brain works as a syntactic engine that recognizes symbol structures and deterministically transforms them to other structures according to algorithms [2, p. 282].

Accuracy vs. Application

When attempting to build a theoretical model of a real-world phenomenon, there is often a tension between accuracy and usefulness. Highly accurate models are often too abstract or computationally intractable to apply in a meaningful way. Conversely, models that see wide application often come with caveats or assumptions that call into question whether the underlying phenomenon truly behaves as the model would predict. The best models carefully balance this negotiation, making justified trade-offs to achieve their intended goal.

Bayesian epistemology sacrifices strict psychological realism by representing what amounts to the result of neurons firing with artificially precise, continuous probabilities; this is a structural idealization employed in ultimate pursuit of the "expectation" in "expected utility theory". Similarly, the LoT hypothesis purposefully abstracts away from the chemistry of the brain in order to model cognition via discrete symbols that behave like a language. This biological inaccuracy is motivated by the LoT's graceful explanation of the infinite productivity and apparent systematicity of human reasoning.

Limitations of the Classic LoT Hypothesis

While the classical LoT Hypothesis is an effective explanation for the productivity of thought (i.e., LoT explains how an infinite number of distinct thoughts can be generated from a finite set of primitives), it cannot adequately explain how minds handle uncertainty. The architecture of the LoT hypothesis is built on discrete and binary representations; the symbols hypothesized by the LoT evaluate strictly to true or false. The syntactic system of a language is structurally unequipped to compute continuous, graded degrees of partial belief and how they are conditionally updated as new information is acquired [1, p. 4]. If a cognitive system recognizes only the syntactic structure of a discrete symbol alone, it is hard to explain how this syntax would be able to encode the fine-grained difference between a 60% degree of belief and a 61% degree of belief without heavy, computationally expensive abstraction. To unify the compositional productivity of the LoT hypothesis with the continuous conditional degrees of partial belief of probabilism, the *Probabilistic Language of Thought hypothesis* modifies the LoT hypothesis to replace deductive logic with generating stochastic processes. This is an important change: the classic LoT Hypothesis relies on deductive rules that preserve the categorical truth of the symbols with certainty, whereas probabilism posits that belief is an inherently inductive process that updates continuously based on the weighted credences of new evidence via Bayes' theorem.

The Probabilistic Language of Thought Hypothesis

In a 2014 paper titled "Concepts in a Probabilistic Language of Thought" Goodman, Tenenbaum and Gerstenberg introduced the Probabilistic Language of Thought Hypothesis hypothesis (PLoT) [1, p. 1]. The PLoT makes the following modifications to the computational theory of mind employed by the classical LoT hypothesis:

The Stochastic λ-Calculus & Generative Processes: Mental representations are *Generative Processes* ("the series of random steps by which the world comes to be as it is" [1, p. 4]). By introducing "primitive randomness" to a deterministic process of discrete compositional steps, it becomes a stochastic process. Rather than resulting in a single definite output, a stochastic process results in a probability distribution over all possible outputs[2]. Since the steps taken by these stochastic processes are each a valid syntactic manipulation, the resulting probability distribution inherits the compositionality that the deterministic λ-calculus provides.

Concepts as Memoized Stochastic Functions: Rather than a concept statically mapping a symbol to a definitive truth value, a concept is a stochastic function, meaning that it applies prob-

[2] More precisely, a stochastic process is indexed by time, and each time index is associated with a probability distribution. Such a process does not necessarily converge to a single stable distribution across time steps, unless the specific process happes to be particularly "well-behaved".

abilistic weights to a symbol [1, p. 7]. To prevent inconsistencies that would arise from evaluating the exact same concept twice and getting two different results, the evaluations of concepts are *memoized*. A memoized stochastic function is evaluated once for a given argument; the result of this single evaluation is then stored in memory [1, p. 6]. Then, whenever the function is called on the same argument in the future, the stored result is used rather than recomputing the function each time it is needed. Memoization ensures the logical stability of concepts, which is necessary for concepts to remain identical to themselves across complex thoughts while still encoding uncertainty in the fundamental process of thought.

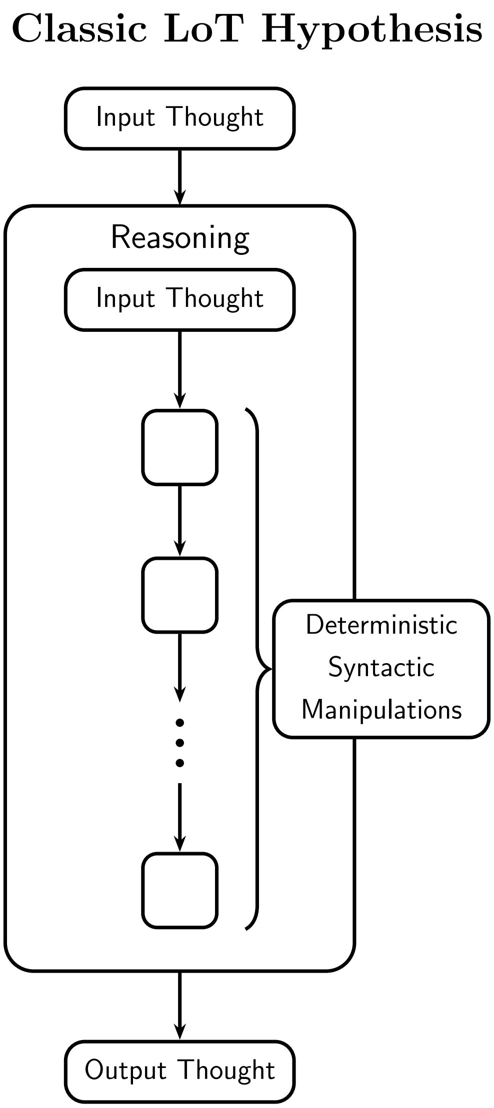
Sampling Semantics: By turning concepts into stochastic functions, mental processes deal with distributions over computations rather than dealing with individual computations. Instead of expecting the brain to explicitly calculate the entirety of these probability distributions, the cognitive system evaluates only a single randomly sampled outcome of each stochastic function being composed [1, p. 6]. This selection of individual computations is known as *Sampling Semantics*. This implies that the composition of probabilities is achieved organically through the ordinary composition of functions.

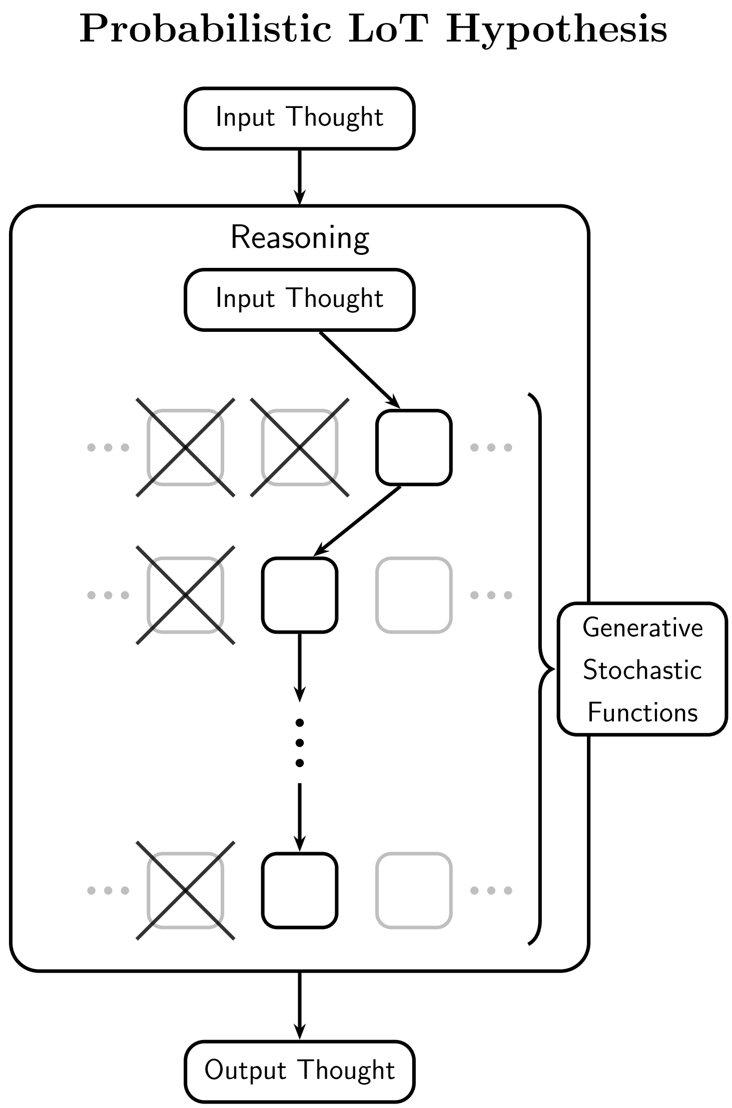
Reasoning as Probabilistic Conditioning (Rejection Sampling): Rather than a mechanical separation and combination of symbols, the fundamental method of updating beliefs is via probabilistic conditioning. Formally, this is modeled via *Rejection Sampling*, in which the mind has a prior distribution (a stochastic generative function) and a specific condition (the new information). To proceed with reasoning, the mind randomly samples the prior distribution repeatedly, and the samples that satisfy the condition form the posterior distribution, while the samples that do not satisfy the condition are rejected [1, p. 6].

Illustrative Example: To make these four modifications concrete, consider an agent Alice forming a belief about tomorrow's weather. Her mental concepts acting as memoized stochastic functions inject primitive randomness into each compositional step of thought rather than mapping deterministically to a static truth value. The generative combination of her concepts defines a higher-order prior probability distribution over a space of possible credence distributions regarding three mutually exclusive candidates: *Clear Skies*, *Light Rain*, and *Thunderstorms*.[3] Via Sampling Semantics, Alice's cognitive system draws a single sample from this higher-order prior. This initial sample is a specific credence distribution, suppose here that it is one heavily weighting *Thunderstorms*, consistent with the severe forecast she reads on her weather app. However, Alice recalls that her app has been wrong before. Rather than accepting this sample outright, her cognitive system in-

[3] For demonstration, this simplified three-outcome space is assumed; in practice, the prior distribution could range over any number of competing hypotheses.

vokes rejection sampling: since the sampled distribution fails to align with her doubt in the app's reliability, it is rejected and a new sample is drawn from the prior generative process. Through this filtering, the system organically converges on an accepted posterior divided between *Light Rain* and *Thunderstorms*, while *Clear Skies* is eliminated entirely. This posterior satisfies the axioms of probabilism and reflects a rational integration of prior belief and new evidence through conditionalization, achieved organically through the composition of stochastic functions.





Classic LoT: Reasoning follows a sequence of deterministic syntactic manipulations. Because the architecture relies on discrete true/false inputs, it evaluates along one definitive sequence, lacking the structural capacity to handle uncertainty or graded degrees of partial belief.

Probabilistic LoT: Concepts operate as generative stochastic functions. Through *Sampling Semantics*, the cognitive system draws specific credence distributions from a higher-order prior (represented by the individual boxes). If a sampled distribution is probabilistically incompatible with current beliefs (e.g., app unreliability), it is eliminated via *Rejection Sampling* (crossed-out boxes).

These modifications preserve the productivity and the compositional nature of the classic LoT Hypothesis while explaining how beliefs are updated probabilistically via a process of conditionalization. This model retains the strengths of the classic LoT Hypothesis and also satisfies the necessary assumption of probabilism.

Limitations of the PLoT Hypothesis

Even if the PLoT unifies the LoT hypothesis and probabilism, it still inherits many of the limitations and caveats of its parent theories. One such generational curse afflicting the PLoT is the so-called "relevance problem," which originates as a limitation of the classical LoT [2, p. 291]. The problem is that it is not clear how the mind determines the scope of relevant information for a given reasoning task. For example, when deciding whether fries or chips are preferred with lunch, the mind knows that recalling the quadratic formula is unnecessary. In a computational system, determining relevance typically involves searching through all stored information to find the relevant pieces and discarding the rest; this seems to demand an enormous amount of computational effort for something that minds do nearly instantly and constantly.

In the PLoT, sampling semantics keep the computational complexity of any single sequence of reasoning tasks identical to the corresponding sequence in the LoT (assuming that sampling a single observation from a generative stochastic function has the same time complexity as a deterministic syntactic manipulation). However, the PLoT is a more computationally demanding theory overall, a consequence of its loftier goals. The classical LoT is built on radical concept nativism, the idea that all primitive concepts are innate and no new primitive concepts can be learned [2, p. 287]. This is in stark contrast to the PLoT, which leverages all of Bayesian epistemology to explain concept acquisition and refinement.

The PLoT employs "probabilistic program induction" as a solution to primitive concept acquisition [1, p. 19]. Because the PLoT represents concepts as stochastic functions, learning a new concept from examples is fundamentally a problem of program induction. This induction is solved via Bayesian inference: a higher-order "program-generating program" produces candidate mental programs, which the cognitive system evaluates by conditioning on observed examples to find the conceptual rule that best explains the evidence [1, p. 19]. Unlike Fodor's nativist solution, this gives the PLoT more explanatory power but at the cost of an order-of-magnitude increase in computational activity; Bayesian induction over an infinite space of possible mental programs generates a massive combinatorial explosion [1, p. 20], which dramatically exacerbates the relevance problem.

Conclusion

The classical LoT hypothesis provides a strong account of the structure and productivity of thought but lacks the architecture to represent and update beliefs under uncertainty. The PLoT addresses this by embedding probabilistic reasoning within a compositional framework, unifying structured representation with graded belief. This comes at the cost of increased computational complexity and an intensified relevance problem, and whether these costs can be reconciled with the constraints of real cognitive systems remains an open question.

References

[1] Goodman, N., Tenenbaum, J., & Gerstenberg, T. (2014). Concepts in a Probabilistic Language of Thought. `https://cocolab.stanford.edu/papers/GoodmanEtAl2015-Chapter.pdf`

[2] Schneider, S. (2009). The language of thought. In S. Robins, J. Symons, & P. Calvo (Eds.), The Routledge companion to philosophy of psychology (pp. 280–295). Routledge. `https://schneiderwebsite.com/uploads/8/3/7/5/83756330/lot_for_the-routledge-companion-to-philosophy-of-psychology.pdf`

[3] Titelbaum, M. G. (2022). Fundamentals of Bayesian Epistemology 1. Oxford University Press.