Towards a more flexible Language of Thought: Bayesian grammar updates after each concept exposure

Pablo Tano\textsuperscript{1}, Sergio Romano\textsuperscript{2,3}, Mariano Sigman\textsuperscript{4}, Alejo Salles\textsuperscript{1}, and Santiago Figueira\textsuperscript{2,3}
\textsuperscript{1}CONICET-Universidad de Buenos Aires. Instituto de Cálculo. Argentina
\textsuperscript{2}CONICET-Universidad de Buenos Aires. Instituto de Ciencias de la Computación. Argentina.
\textsuperscript{3}Universidad de Buenos Aires. Facultad de Ciencias Exactas y Naturales. Argentina
\textsuperscript{4}Universidad Torcuato di Tella. Argentina.

Recent approaches to human concept learning have successfully combined the power of symbolic, infinitely productive rule systems and statistical learning to explain our ability to learn new concepts from just a few examples. The aim of most of these studies is to reveal the underlying language structuring these representations and providing a general substrate for thought. However, describing a model of thought that is fixed once trained is against the extensive literature that shows how experience shapes concept learning. Here, we ask about the plasticity of these symbolic descriptive languages. We perform a concept learning experiment that demonstrates that humans can change very rapidly the repertoire of symbols they use to identify concepts, by compiling expressions which are frequently used into new symbols of the language. The pattern of concept learning times is accurately described by a Bayesian agent that rationally updates the probability of compiling a new expression according to how useful it has been to compress concepts so far. By portraying the Language of Thought as a flexible system of rules, we also highlight the difficulties to pin it down empirically.

I. INTRODUCTION

How can children acquire a vast universe of concepts with seemingly very little exposure? One possible solution to this conundrum, known as the Plato Problem \cite{1,2}, builds on the human capacity to describe concepts—and more generally of all elements of thought—through the use of a symbolic and combinatorial mental language \cite{3}, referred as language of thought (LoT) \cite{4}.

Combinatorial languages can describe a vast set of concepts from a small set of primitives. This can be understood in a relatively simple example in the domain of shapes. A combinatorial and symbolic language similar to Logo \cite{5} can combine operations such as “move”, “pen up”, “pen down” or “rotate” to generate an infinite set of expressions (or programs) which, when evaluated, can convey all sorts of shapes.

A language describing concepts (like shapes) also provides a natural notion of their complexity \cite{6}. A concept is simple, relative to that language, when it can be described by a short program. On the contrary, it is complex when all its descriptions require a long sequence of instructions. For example, in the case of the Logo language, a square can simply be instructed as a loop of four displacements followed by rotations of 90 degrees. In this language, the icon of a face will be constructed as a loop of four displacements followed by rotations of 90 degrees. A concept is simple, relative to that language, when it can be described by a short program. On the contrary, it is complex when all its descriptions require a long sequence of instructions. For example, in the case of the Logo language, a square can simply be instructed as a loop of four displacements followed by rotations of 90 degrees. In this language, the icon of a face will be constructed as a loop of four displacements followed by rotations of 90 degrees.

In the example of the Logo language one can imagine that if productions which draw squares are very frequent, it would be efficient to devote a new symbol to this production. The new symbol ‘square’ is a hierarchical ‘second order’ construction of the ‘first order’ primitives of the language. It has a cost (of increasing the lexicon of the language) but in the new language, drawing a square can be instantiated with a very short program (namely, ‘square’) and hence uses less memory.
Indeed, a higher level language allows us to reach a higher level of abstraction by freeing memory and processing power, thus making more complex thoughts thinkable [13][14].

Most work in the LoT literature, while naturally including a learning mechanism, tends to approach the LoT as a stable system to be unearthed by experimenters, who try different candidate templates and select the one which best fits the data after training [8][13][14]. Still, how different tracks of experience can shape acquisition differently and can constantly change the repertoire of a LoT after each exposure remains to be discovered.

Here, we perform a Boolean concept learning experiment to show that humans can change very rapidly—in the course of an experiment—the repertoire of symbols they use to identify concepts. We also provide a dynamic model that is flexible enough to update its underlying language after each concept exposure.

In our experiment, participants are divided in two groups, in such a way that each group is presented with a different sequence of concepts. One of the two groups is presented with concepts that are succinctly described only if the logical operator ‘exclusive or’ (xor, notated $\oplus$) is used, which we presume does not form part of the natural repertoire of LoT in this specific domain [8]. However, these concepts can also be described with a sensibly lengthy combination of primitives excluding $\oplus$. We show how the exposure to this set of concepts ‘compiles’ the operator in a way that, after exposure, subjective difficulty is described by an extended language in which $\oplus$ has been incorporated to the set of primitives. Furthermore, we show that the subjective difficulty of concepts throughout the task is consistent with that of a Bayesian agent that rationally updates the probability of compiling $\oplus$ according to how useful it has been to compress concepts so far.

II. THE LOGICAL SETTING

We consider two propositional logics, both containing only four propositional variables $\text{Vars} = \{x_1, x_2, x_3, x_4\}$. $\text{P}$ is defined over the signature $\land, \lor$ and $\neg$, and $\text{P}^\oplus$ is defined over the signature $\land, \lor, \neg$ and $\oplus$. As one can see from the grammars defined in Fig. 1 the only difference between $\text{P}$ and $\text{P}^\oplus$ is that the latter has an additional operator $\oplus$.

**Fig. 1**: The context free grammar for language P. Language $\text{P}^\oplus$ has an extra rule: $\text{BOOL} \rightarrow \text{BOOL} \land \text{BOOL}$

The semantics of $\land, \lor$ and $\neg$ are standard: conjunction, disjunction and negation, respectively. We let $\oplus$ denote the exclusive disjunction. As usual, $v \models \varphi$, represents that the formula $\varphi$ is true for the valuation $v : \text{Vars} \rightarrow \{0, 1\}$ and we denote the semantics of $\varphi$ by $\llbracket \varphi \rrbracket = \{v : v \models \varphi\}$. A concept $C$ is a set of valuations $\text{Vars} \rightarrow \{0, 1\}$. The complement of $C$ is denoted $\overline{C}$ and is defined as $\overline{C} = \{0, 1\}^{\text{Vars}} \setminus C$. Observe that $\#C + \#\overline{C} = 16$. We say that a formula $\varphi$ is compatible with concept $C$ if $\llbracket \varphi \rrbracket = C$. We regard logics as languages for describing concepts. Any concept $C$ has infinitely many descriptions, namely, all formulas $\varphi$ such that $\llbracket \varphi \rrbracket = C$.

Example. In Fig. 2 we depict a concept $C$ (variables are represented by colors) such that $\#C = 4$. One can see that the formula $x_3$ is not compatible with $C$ but $x_1 \land x_2$, or $x_1 \land x_2 \land (x_4 \lor \neg x_3)$, are compatible with $C$. $\overline{C}$ may be described by $\neg x_1 \lor \neg x_2$.

We will often identify concepts with any formula compatible with it, so we will talk of “concept $\varphi$” to refer to “concept $\llbracket \varphi \rrbracket$”. However, it should be noted that a concept is a semantic object that has many descriptions in the logical language.

Our interest is in the size of the shortest descriptions of a given concept, that is, in the minimum description length (MDL) of concepts.

The size of a formula $\varphi$ is denoted $|\varphi|$ and it is defined as the number of operators plus the number of atoms (i.e. possibly negated propositional symbols), that is: $|x_i| = |\neg x_i| = 1$ for $i = 1 \ldots 4$ and $|\varphi_1 \lor \varphi_2| = |\varphi_1| + |\varphi_2| + 1$ for $* \in \{\land, \lor, \oplus\}$. For $L \in \{P, P^\oplus\}$ and a concept $C$ we define the minimum description length of $C$ with respect to $L$ as

$$\text{MDL}_L(C) = \min\{|\varphi| : \varphi \in C, \llbracket \varphi \rrbracket = C\}.$$  

Since $P$ is a sublanguage of $P^\oplus$, we have $\text{MDL}_{P^\oplus}(C) \leq \text{MDL}_P(C)$ for any concept $C$.

Example. The concept $C = \{v : v(x_1) + v(x_2) = 1\}$, which expresses that $x_1$ is true or $x_2$ is true but not both can be described in $P^\oplus$ as $\varphi = x_1 \oplus x_2$, of length 3. In fact, one can check that this is the shortest formula compatible with $C$, and so $\text{MDL}_{P^\oplus}(C) = 3$. If we now switch to $P$, we can no longer describe $C$ as $x_1 \oplus x_2$, since $\oplus$ is not part of its signature. However, in $P$, the concept $C$ may be described by formula $\psi = (x_1 \land \neg x_2) \lor (x_2 \land \neg x_1)$, of size 7. Since this formula has minimal size, we have that $\text{MDL}_P(C) = 7$.

III. EXPERIMENT

55 participants participated in the experiment over the world wide web using the Amazon Mechanical Turk crowd sourcing platform. All were US residents over the age of 18 and had more than 95% of past tasks successfully approved by other requesters. 44 participants completed the experiment through all the stages and declared not cheating (using pen, screenshots or a similar method to copy the answers) at the end of the
As shown in Table I, we presented the target group with training concepts which are succinctly described when $\oplus$ is part of the language, but necessarily described with lengthier formulas if $\oplus$ is absent. More technically, concepts for which $\text{MDL}_{p\oplus}$ is much smaller than $\text{MDL}_p$.

Participants in the control group, on the other hand, experienced a sequence of concepts that could be easily described using the language given by $P$. After these training concepts, both groups were presented with the same pair of test concepts: one which could be only succinctly described in $P^{\oplus}$, and one for which the MDL did not depend on the underlying language $P^{\oplus}$ or $P$. We compared learning times between the two groups for these last two concepts.

As shown in Table II, training concepts for the target (xor) group were $x_1, x_1 \oplus x_j, x_1 \oplus x_j \oplus x_k,$ and $x_k \oplus x_l$, called $C^1, C^2, C^3, C^4$ respectively. Training concepts for the control group were $x_1, x_1 \lor x_j, x_1 \lor (x_j \land x_k),$ and $x_k \lor x_l$ called $C^1, C^2, C^3, C^4$ respectively. We use the indexes $i, j, k, l$ instead of numbers because variables were randomized in each trial. $x_1$ could stand for $x_1, x_2, x_3$ or $x_4$, that is, for any of the four colors. After these four concepts, both groups were presented with the same test concepts: $x_1 \land (x_j \oplus x_k),$ and $x_1 \land (x_j \lor x_k)$, called $C^5$ and $C^6$ respectively.

Choosing which concepts to show the target group in order for them to ‘learn’ the $\oplus$ operator is critical in our experiment. Crucially, the learner must have an option between two alternatives that describe the concept: one that is succinct but uses $\oplus$, or necessarily a much longer one in the absence of $\oplus$. In other words, these concepts must be compatible with short logical formulas if and only if we take $P^{\oplus}$ as the language of description. To ensure that this was the case, we enumerated, for each concept, all formulas compatible with it and produced by the $P$ and $P^{\oplus}$ grammars up to length 19. For all training concepts of the target group, the shortest compatible formula without $\oplus$ is much longer than the shortest compatible formula with $\oplus$. This is shown in Table II.

### IV. MODEL-FREE RESULTS

We measure the subjective difficulty of a given concept as the total time needed by the participant to successfully encode the concept, which indicates that they can reliably express which exemplars belong to the concept and which do not.

Participants from the target group spent almost half the time participants from the control group spent on the same concepts, which could be succinctly described only in $P^{\oplus}$ $(111 \pm 16$ s.e.m. seconds versus $214 \pm 37$ s.e.m. seconds, a two-sample t-test reveals...
which each rule (symbols and operators) is associated with

throughout the concept sequence (the 6 concepts in Table I).

has a very low probability of being used, formulas that require
the rules required for building it, and therefore it decreases

generate formulas, the model uses a symbolic language in

generates logical formulas and evaluate them to true or false

get group represented the

after exposure to the training concepts, participants in the tar-

subjective difficulty of learning a concept is proportional to

cured
C
5
= d
C
5

5
2
6

T
42
= 2.6, P < 0.01), as shown in Fig.3(left). We also found
that the control group learned much faster C
5
(143 ± 14 s.e.m.
seconds for the target group versus 76 ± 10 s.e.m. seconds for
the control group, T
42
= 3.5, P < 0.01). A mixed ANOVA
with C
5
-C
6
as within subject factor and target-control groups
as between subject factor reveals a strong interaction between
group and C
5
-C
6
(F = 15.3, P < 0.001), indicating that the
differences in learning times for C
5
and C
6
were very different
between the two groups.

The target group encoded C
5
more efficiently than the con-
trol group. We propose that the control group expected to find
in C
5
and C
6
structures that could be easily built in P. The
target group, on the other hand, became biased towards the
⊕
structure, and they expected to find it in C
5
and C
6
. This
cased C
5
to be encoded more rapidly by the target group and
C
6
more rapidly by the control group. Assuming that the
subjective difficulty of learning a concept is proportional to
the complexity of its internal representation, we conclude that
after exposure to the training concepts, participants in the tar-
group represented the ⊕ more efficiently than the control
group, and expected to find this structure in C
5
and C
6
.

V. MODEL

When presented with a concept (e.g. Fig. 2), our model
generates logical formulas and evaluate them to true or false
for that concept, keeping the formula only if it is true. To
generate formulas, the model uses a symbolic language in
which each rule (symbols and operators) is associated with
a probability of being used. The probability of generating a
formula is proportional to the product of the probabilities of
the rules required for building it, and therefore it decreases
exponentially with its length. Furthermore, if one of the rules
has a very low probability of being used, formulas that require
it will also have very low probability.

The Static model maintains the rules’ probabilities fixed
throughout the concept sequence (the 6 concepts in Table[1]).

The Dynamic model updates the probabilities after each con-
cept, in order to minimize the expected description length of
future concepts, assuming they have similar structure to the
concepts learnt so far. We include in this model the ⊕ rule a
priori in the language, but with vanishing probability of being
used. Changes in this probability can be analogously inter-
preted as the probability that a rational agent without the com-
plied symbol a priori decides to add the compiled expression
as a new primitive into her language.

A. Static Model

Under the LoT assumption, given a concept C (e.g. Fig.2),
the probability that an agent uses formula \( \varphi \) to explain this
concept is defined by Bayes theorem:

\[
P(\varphi \mid C) \propto P(C \mid \varphi)P(\varphi).\]

The likelihood \( P(C \mid \varphi) \) of a logical statement \( \varphi \) can be simply
defined as 1 if \( \|\varphi\| = C \) and 0 otherwise. In other words, for
given any concept, only explanations that describe this concept
are considered as possible explanations. The likelihood term
has been defined more flexibly in the literature [8,15], allowing
for mislabeled elements. We keep this simpler definition in
order to reduce the number of free parameters of the model,
as we do not intend to account for mislabeling errors in our
experiment.

The prior \( P(\varphi) \) is defined by augmenting the context-free
grammars shown in Fig.1 into a probabilistic context-free
grammars (PCFG). In the PCFG, each rule has associated a
parameter indicating the probability of using that rule. A PCFG
can be used to produce logical statements similar to a CFG.
Each non-terminal remaining in the statement is expanded
using a rule of the PCFG with probability proportional to that
rule’s associated parameter, until no non-terminals remain in
the statement.

We assume that the probability that a subject uses formula \( \varphi \)
to explain concept C is proportional to the posterior \( P(\varphi \mid C) \),
and the subjective difficulty \( d_C \) of a concept C to a participant
is proportional to the length of the formula that the participant
is using to explain that concept. However, there is no way to know
directly which internal formula \( \varphi \) the participant is using (and
therefore we do not know \( |\varphi| \)). Hence, the most parsimonious
approach is to consider the entire posterior distribution \( P(\varphi \mid C) \)
over possible formulas[2].

Given a concept C, the expected length \( E_C \) of the formulas
used by the participant is simply

\[
E_C = \sum_{|\varphi|=C} |\varphi| P(\varphi \mid C),
\]

2 This is equivalent to the Sampling Hypothesis described in [17], by which
participants represent distributions through samples. Similar results are
obtained if each participant carries entire probability distributions.
where the sum is over all formulas $\varphi$ compatible with $C$. We define the difficulty $d_C$ of a concept experienced by the participant as

$$d_C \propto E_C + \alpha N_C,$$

where we added a term that accounts for the cardinality of the concept: $N_C$ is the cardinality of the concept or its complement, the one being smaller, i.e. $N_C = \min \{ |C|, \#C \}$ (e.g. $N_C = 4$ for the concept $C$ of Fig. [2]), and $\alpha$ is a free parameter fitted globally for all concepts and participants to its maximum likelihood value of 0.9. In this way, we remove the asymmetry between positive and negative examples, while accounting for the toil taken by considering a larger number of items simultaneously.

In practice, to approximate $E_C$ for each concept $C$, we calculated the posterior probability $P(\varphi \mid C)$ of all compatible formulas $\varphi$s up to size 19 with $P(\varphi \mid C)$ and then use (1). Since the space of all possible $\varphi$s grows exponentially with $|\varphi|$, normative procedures for estimating $P(\varphi \mid C)$ in this space involve stochastic search algorithms. However, in our case, we were able to exhaustively enumerate and calculate the posterior probability of all formulas generated by the PCFG up to a sufficiently high size $M$ such that all formulas with $|\varphi| > M$ have vanishing probabilities when compared to shorter compatible formulas for the current concept (because the prior $P(\varphi)$ decreases exponentially with the size of the formula).

### B. Dynamic Model

Up to this point, we assumed that, given a concept $C$, the posterior distribution over formulas $P(\varphi \mid C)$ was independent of the other concepts presented in the sequence. However, if the LoT (i.e. the PCFG) updates with experience, the prior $P(\varphi)$ in $P(\varphi \mid C)$ will change, and so will $E_C$ in (1) and finally the subjective difficulty $d_C$. Therefore, $d_C$ will depend on the sequence of concepts that were previously presented to the participant.

In other words, since now $P(\varphi)$ depends on the sequence of concepts experienced by the participant, instead of $P(\varphi \mid C)$, we have

$$P(\varphi \mid C^t, \ldots, C^1) \propto P(C^t \mid \varphi) P(\varphi \mid C^1, \ldots, C^{t-1})$$

, where $C^t$ is the concept presented at trial $t$, and $P(\varphi \mid C^1, \ldots, C^{t-1})$ depends on the state of the PCFG at trial $t$, which in turn depends on how the PCFG gets updated from trial to trial.

Intuitively, the update process increases the probability of using a certain rule in the PCFG accordingly to how useful this rule was to compress compatible formulas for the concepts previously learned in the same domain. Specifically, we model the update process in a normative manner: the probability of using a rule of the PCFG at trial $t$ is equal to the Bayesian posterior probability that this rule will enable the learner to find compressed explanations at trial $t$, according to how useful it was to compress explanations in trials $1, \ldots, t-1$.

To formalize the update of the PCFG, we define $P(\varphi)$ similarly to [5]. Specifically, the prior probability of a logical statement at trial $t$ in the concept sequence uses a single Dirichlet-multinomial for the set of rule expansions. The Dirichlet is parameterized by a set of positive real numbers $D^i$, one for each rule $i$ in the PCFG, which in turn determine the probability of using rule $i$ at trial $t$: a higher $D_i$ indicates a higher probability of using rule $i$.

The prior is specified by the set Dirichlet parameters $D^0$ with which we start the experiment ($D^0$ represents a vector containing the prior parameters of all rules in the grammar at trial 0). In our experiment, we set the prior Dirichlet parameters of all rules equal to 1, and the parameter of the rule that expands the target operator to a value several orders of magnitude smaller ($= 10^{-4}$). This means that the target operator was practically absent at the beginning of the experiment, but it was technically possible to ‘learn it’ by increasing its probability as the experiment developed.

Under the Dirichlet model, the prior $P(\varphi \mid C^1, \ldots, C^{t-1})$ can be rewritten using the Dirichlet parameters as $P(\varphi \mid D^t)$. Therefore, to know how $P(\varphi \mid C)$ updates from trial to trial, we only need to know how $D$ updates from trial to trial.

The Dirichlet parameter of rule $i$ at trial $t$ + 1 is equal to its parameter at trial $t$ plus the amount of times the production $i$ was used in generating all formulas compatible with the concept at trial $t$ (we note $M_i(\varphi)$ as the number of times that rule $i$ is used in generating formula $\varphi$), weighted by each formula’s posterior probability at trial $t$:

$$D_i^{t+1} = D_i^t + \sum_{[\varphi]=C^t} P(\varphi \mid D^t) M_i(\varphi). \quad (2)$$

This Bayesian learning mechanism increases the probability of using rules that allow concepts to be succinctly described. This happens because these formulas have higher probability $P(\varphi \mid D)$ than longer formulas, so the Dirichlet parameters of the rules that build these formulas increase more strongly than those of the rules that build longer formulas.

### VI. RESULTS

The Bayesian agent that minimizes the expected complexity of future concepts by optimally adapting its LoT to the inferred structure of the task accurately captures the dynamics of human learning across concepts. If we did not allow the model to update the probability of the operators after each concept, and particularly the compiled operator $\oplus$, the control group and the target group would be indistinguishable to the model as it would predict equal average formula length for both groups (see Fig. [3] Static Model). Instead, as shown in Fig. [4] by adjusting the prior probabilities based on concept exposure the dynamic model is able to capture learning time patterns in the target groups ($R^2 = 0.96$ compared to $R^2 = 0.73$ for the static model). Expectedly, both models perform similarly in the control groups as they were designed to not encourage the use of any particular operator ($R^2 = 0.72$; $R^2 = 0.71$ for the static model). The impact of the learning capability of the model is
most evident in the target group concept sequence, which was
designed to this effect. If the structure of the concepts does
not bias the LoT primitives one way or the other, it is expected
that a static model will provide a reasonable fit. However, it is
difficult to tell a priori how unbiased a set of concepts really is,
so experiments relying on repeated concept exposure should
always take between-concept learning into account.

Allowing the model to constantly update its beliefs from
concept to concept is a requisite to capture human learning
times. We now explain how the pattern of subjective difficulties
in Fig. 4 emerged in the Dynamic model. In this scenario,
learning for the model is formalized by the update of rule
parameters from concept to concept according to (2).

In Fig. 5 we show how this learning takes place in the concept
sequence for the target group. There are mainly two competing
formulas when $C_t^2$ is presented: $x_i \oplus x_j$ and $(x_i \land \neg x_j) \lor (\neg x_i \land x_j)$. Given the low a priori value of the parameter of the $\oplus$ rule,
the posterior of the formulas of type $(x_i \land \neg x_j) \lor (\neg x_i \land x_j)$,
which do not use the $\oplus$ operator, is much higher than the
posterior of $x_i \oplus x_j$. Therefore, in Fig. 4 (middle top) we see a
large predicted difficulty by the dynamic model for this concept
(since the posterior lies mainly over these longer formulas
without $\oplus$, see [1]).

However, the little increment in the $\oplus$ rule after $C_t^2$ (see
Fig. 5) is sufficient for making the formula $x_i \oplus x_j$ to have
higher relative posterior in the next concepts, making the in-
crement in the parameter of the $\oplus$ rule much greater than
before. Additionally, the difficulty inferred by the model is
much smaller the second time the concept is presented (com-
pare $C_t^1$ and $C_t^2$ concepts in Fig. 4), since now the posterior is
more evenly distributed between long (without $\oplus$) and short
(with $\oplus$) formulas (see [1]). Finally, when the concept $C_t^5$ is
presented, the learner has completely compiled the $\oplus$ rule into
her language, ascribing the formulas that use the $\oplus$ operator a
much higher posterior probability relative to the long formulas
that do not use the $\oplus$ operator. Therefore, the inferred difficulty
for $C_t^5$ is much smaller than those describing previous concepts,
almost as simple as concept $C_t^1$ (see Fig. 4).

VII. DISCUSSION

We measured the subjective difficulty that participants expe-
xience when learning a sequence of concepts. To explain this
subjective difficulty, we resource to propositional logic as a
base description language. In the target group we experimented
with concepts which can be succinctly described in the base
language that also contains an extra operator $\oplus$ for exclusive
disjunction but that needed necessarily longer descriptions over
the base language (where this operator is absent). On the con-
trary, the control group is exposed to concepts where $\oplus$ does
not help to achieve succinctness.

Learning times are consistent with the hypothesis that par-
ticipants in the target group smoothly adopt the $\oplus$ as a new
primitive of their LoT in order to absorb the concepts they
have been exposed to, with no more incentive than decreasing
the expected complexity of future concepts. We do not claim
that participants have learned the $\oplus$ operator defined by any
specific formula using the previous operators, however, their
LoT seems to have constructed an operation that matches the
semantics of the exclusive or in order to compress such patterns
of data and identify them more efficiently.

Here, we focus on transfer learning effects when learning
sequential concepts that share the same hierarchical structure.
We acknowledge, however, that several other transfer learning
effects are present in human sequential logical concept learn-
ing, such as when subsequent concepts differ in the relevant
variables (e.g. color lights in our experiment) [18], when chang-
ing the relevant variables in subsequent exclusive disjunctions
[19], or when two categories are learned in an interleaved or a
focused manner [20]. However, unlike superficial knowledge
about the task (like the frequency of appearance of different
symbols and logical operators in the concept sequence),
identifying the latent hierarchical structure of concepts have
extremely important computational consequences: it allows for
exponentially less complex representations [21][22], maximiz-
ing the expected value of future computations within resource-
bound constraints [23]. In our task, in order to focus primarily
on the learning process of the $\oplus$ structure, we randomize

![Fig. 5: Evolution of Dirichlet parameters of different rules after each concept experienced by the target group.](image-url)
variables in each trial, such that other kinds of transitions are averaged out across participants.

Most LoT studies provide a language that is fixed once trained or inferred over a specific data. We claim that when a specific language beats a second one at fitting some experimental data, what we may be seeing is an effect of prior experience (including from the experiment itself), more than an intrinsic feature of the LoT. This leads to a fundamental difficulty in trying to experimentally uncover what the actual human symbolic substrate of thought is. Experimental results have shown for instance that a grammar with and, or, and not better explains Boolean concept learning than one with nand, despite both being expressively equivalent [8]. In our view, this cannot be taken to mean anything more than that in the current state of affairs of the world, the nand operator is not very useful for compressing information. We have shown that participants can rapidly compile new expressions in their LoT if they begin to be useful, which emphasizes that one cannot simply ignore the order in which concepts are presented to the participant when studying aspects of the LoT.

When Fodor proposed the Language of Thought hypothesis [4], what he had in mind was a symbolic system we all came equipped with from birth. Stating that this language is in fact always flexible might seem in outright contradiction with Fodor’s original idea. In fact, what studies in the LoT literature (including this one) are probably probing is one among many languages in a hierarchy of increasing abstraction. As we progress in life, we find some conceptual summaries useful, and compiled them in a more abstract token. It is even likely that there is no proper hierarchy with sharply defined boundaries between levels, but instead a less organized progression of concepts of increasing abstraction, with thought progressing seamlessly using constructs at different levels.

VIII. CONCLUSION

We defined a model to measure the subjective difficulty of learning a sequence of concepts. The model updates the grammar production probabilities between concepts and predicts difficulty as the size of compatible formulas weighted by their posterior probability. This learning mechanism allows to simulate the emergence of a new primitive in the language, as it becomes useful to encode the concepts presented so far. The predicted difficulties strongly resembles the pattern of human learning times in a sequence of concepts that required the ⊕ operator in order to be efficiently represented.

[1] N. Chomsky, *Knowledge of language: Its nature, origin, and use* (Greenwood Publishing Group, 1986).
[2] N. Chomsky et al., *On cognitive structures and their development: A reply to piaget, Philosophy of mind: Classical problems/contemporary issues*, 751 (2006).
[3] A. Newell, *Physical symbol systems*, Cognitive science 4, 135 (1980).
[4] J. A. Fodor, *The language of thought*, Vol. 5 (Harvard University Press, 1975).
[5] H. Abelson, N. Goodman, and L. Rudolph, *Logo manual*, (1974).
[6] A. N. Kolmogorov, *Three approaches to the quantitative definition of information*, International journal of computer mathematics 2, 157 (1968).
[7] J. Feldman, *The simplicity principle in human concept learning*, Current Directions in Psychological Science 12, 227 (2003).
[8] S. T. Piantadosi, J. B. Tenenbaum, and N. D. Goodman, *The logical primitives of thought: Empirical foundations for compositional cognitive models*, Psychological review 123, 392 (2016).
[9] E. L. Leeuwenberg, *A perceptual coding language for visual and auditory patterns*, The American Journal of Psychology, 307 (1971).
[10] M. Almairic, L. Wang, P. Pica, S. Figueira, M. Sigman, and S. Dehaene, *The language of geometry: Fast comprehension of geometrical primitives and rules in human adults and preschoolers*, PLoS computational biology 13, e1005273 (2017).
[11] S. Romano, A. Salles, M. Almairic, S. Dehaene, M. Sigman, and S. Figueira, *Bayesian validation of grammar productions for the language of thought*, PLoS ONE 13, 1(2018)
[12] G. Lupyan, D. H. Rakison, and J. L. McClelland, *Language is not just for talking: Redundant labels facilitate learning of novel categories*, Psychological science 18, 1077 (2007).
[13] M. L. Minsky, *Computation: finite and infinite machines* (Prentice-Hall, Inc., 1967).
[14] G. L. Murphy, *Comprehending complex concepts*, Cognitive science 12, 529 (1988).
[15] N. D. Goodman, J. B. Tenenbaum, J. Feldman, and T. L. Griffiths, *A rational analysis of rule-based concept learning*, Cognitive science 32, 108 (2008).
[16] C. Kemp, *Exploring the conceptual universe*, Psychological review 119, 685 (2012).
[17] S. Denison, E. Bonawitz, A. Gopnik, and T. L. Griffiths, *Rational variability in children’s causal inferences: The sampling hypothesis*, Cognition 126, 285 (2013).
[18] M. R. Blair, M. R. Watson, R. C. Walsh, and F. Maj, *Extremely selective attention: Eye-tracking studies of the dynamic allocation of attention to stimulus features in categorization*, Journal of Experimental Psychology: Learning, Memory, and Cognition 35, 1196 (2009).
[19] J. K. Kruschke, *Dimensional relevance shifts in category learning*, Connection Science 8, 225 (1996).
[20] P. F. Carvalho and R. L. Goldstone, *Putting category learning in order: Category structure and temporal arrangement affect the benefit of interleaved over blocked study*, Memory & cognition 42, 481 (2014).
[21] Y. Bengio, A. Courville, and P. Vincent, *Representation learning: A review and new perspectives*, IEEE transactions on pattern analysis and machine intelligence 35, 1798 (2013).
[22] B. M. Lake, R. Salakhutdinov, and J. B. Tenenbaum, *Human-level concept learning through probabilistic program induction*, Science 350, 1332 (2015).
[23] S. J. Gershman, E. J. Horvitz, and J. B. Tenenbaum, *Computational rationality: A converging paradigm for intelligence in brains, minds, and machines*, Science 349, 273 (2015).