ABSTRACTING (IN) THE LANGUAGE OF THOUGHT

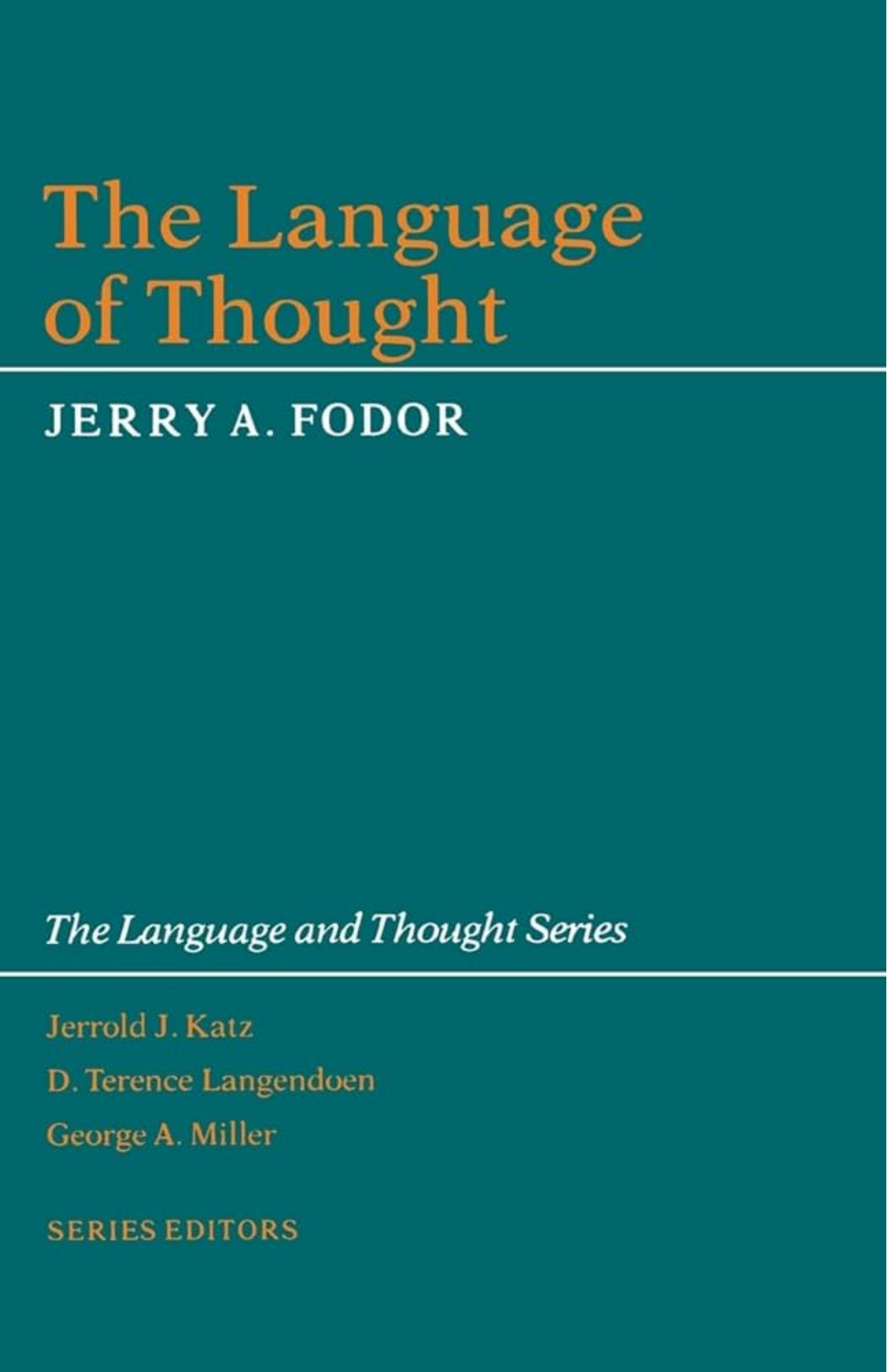
Jakub Szymanik

UNIVERSITÀ DI TRENTO

CiMeC Center for Mind/Brain Sciences

> At the heart of cognitive science is an embarrassing truth: we do not know what mental representations are like.

-Steven Piantadosi, 2020

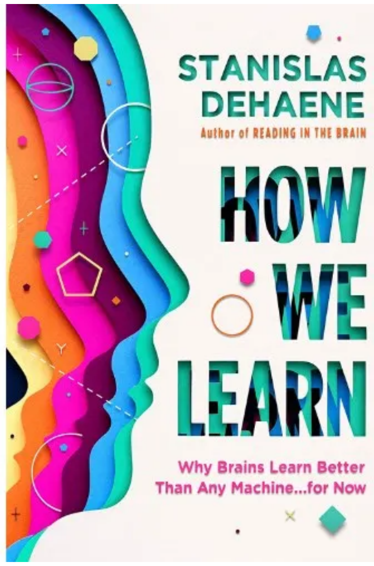


LOT HYPOTHESIS

- The Language of Thought Hypothesis posits that abstraction occurs in a mental language, known as the Language of Thought.
- Generally assumed to look like logic: predicates get combined with logical operators.
- Modern version popularised by Jerry Fodor in the 70s.
- The classical picture of LOT that philosophers have developed is intended to be an account of thinking, explaining phenomena such as learning from a few examples, decision-making, and perception, among others.

WHY ARE WE TALKING ABOUT IT TODAY?

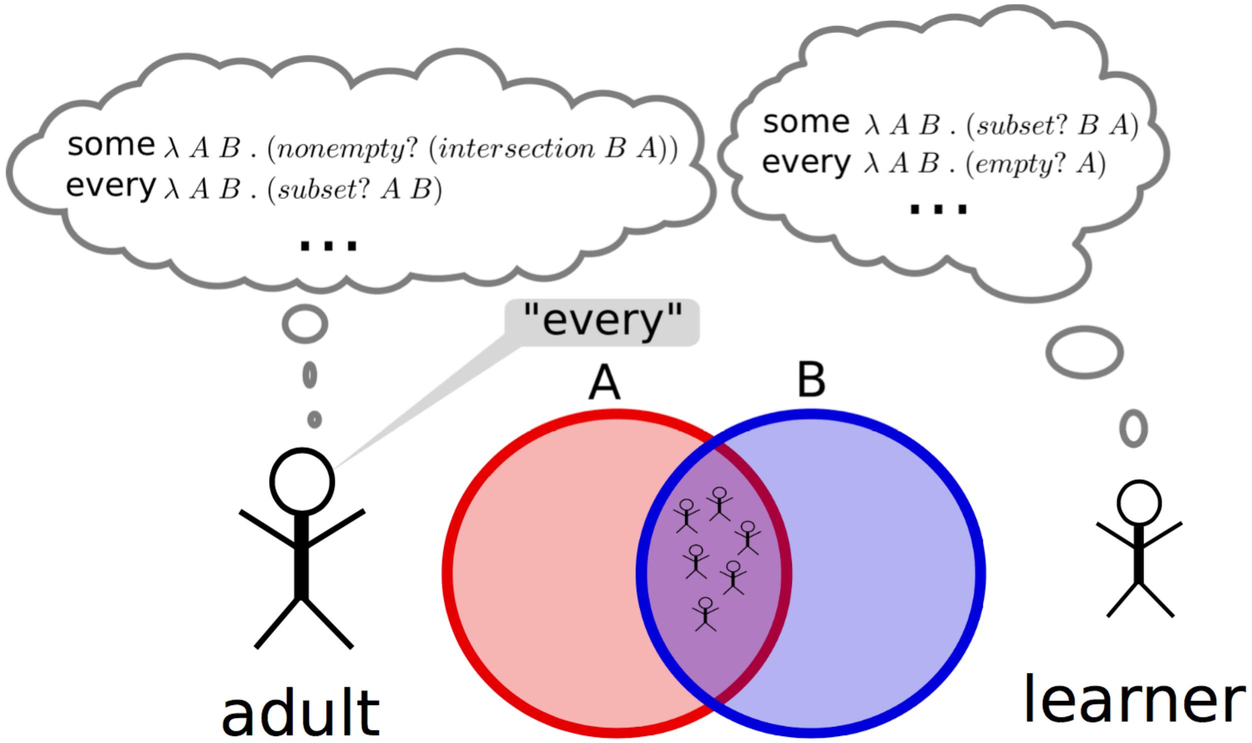
- Leibniz, 1677, Boole, 1854, Fodor, 1975,…, Rescorla, 2019, and others.
- Recently, revived interest in cognitive science:
- Feldman, 2003, Tenenbaum and Griffiths, 2001, Tenenbaum and Xu, 2007, Piantadosi, 2016, Sauerland et al., 2025, Dehaene et al. 2025, etc.
- Perhaps a missing peace to make AI more human-like.
- Current LLMs excel at pattern recognition and statistical inference, but they do not necessarily possess the same symbolic reasoning capabilities as humans.
- True intelligence might require a deeper understanding of the world, potentially through a language-like system for representing and manipulating concepts.



MODERN ***LOT***

- We focus on some small but critical conceptual domains.
- We have an agent with a fairly natural LoT, consisting of primitive concepts and a small set of operators and composition rules.
- We assume people have a simplicity prior: concepts that are harder to express with the LoT have lower prior probability.
- For example,…

EXPLAINING UNIVERSAL CROSS-LINGUISTIC PROPERTIES



Nonterminal		Expansion	Gloss
START	→	$\lambda\ A\ B\ .\ BOOL$	Function of A and B
BOOL	→	*true*	Always true
	→	*false*	Always false
	→	(*card>* *SET SET*)	Compare cardinalities (>)
	→	(*card=* *SET SET*)	Check if cardinalities are equal
	→	(*subset?* *SET SET*)	Is a subset?
	→	(*empty?* *SET*)	Is a set empty?
	→	(*nonempty?* *SET*)	Is a set not empty?
	→	(*exhaustive?* *SET*)	Is the set the entire set in the context?
	→	(*singleton?* *SET*)	Contains 1 element?
	→	(*doubleton?* *SET*)	Contains 2 elements?
	→	(*tripleton?* *SET*)	Contains 3 elements?
SET	→	(*union* *SET SET*)	Union of sets
	→	(*intersection* *SET SET*)	Intersection of sets
	→	(*set-difference* *SET SET*)	Difference of sets
	→	*A*	Argument *A*
	→	*B*	Argument *B*

Piantadosi, Tenenbaum, Goodman. Modeling the acquisition of quantifier semantics : a case study in function word learnability, 2012

LEARNABILITY

- Prior: specifying the learner's estimate of how likely any hypothesis is before any labeled objects have been observed.
- The prior is constructed using the LoT.
- Likelihood: quantifying the probability that the particular set was particularly labeled, if h were the true concept.
- Inferential statistical model: P(h | observed sets and labels).

$$P(m \mid u_1,\ldots,u_n,c_1,\ldots,c_n) \propto P(u_1,\ldots,u_n \mid m,c_1,\ldots,c_n)\,P(m).$$

$$P(u_1,\ldots,u_n \mid m,c_i,\ldots,c_n) = \prod_{i=1}^{n} P(u_i \mid m,c_i).$$

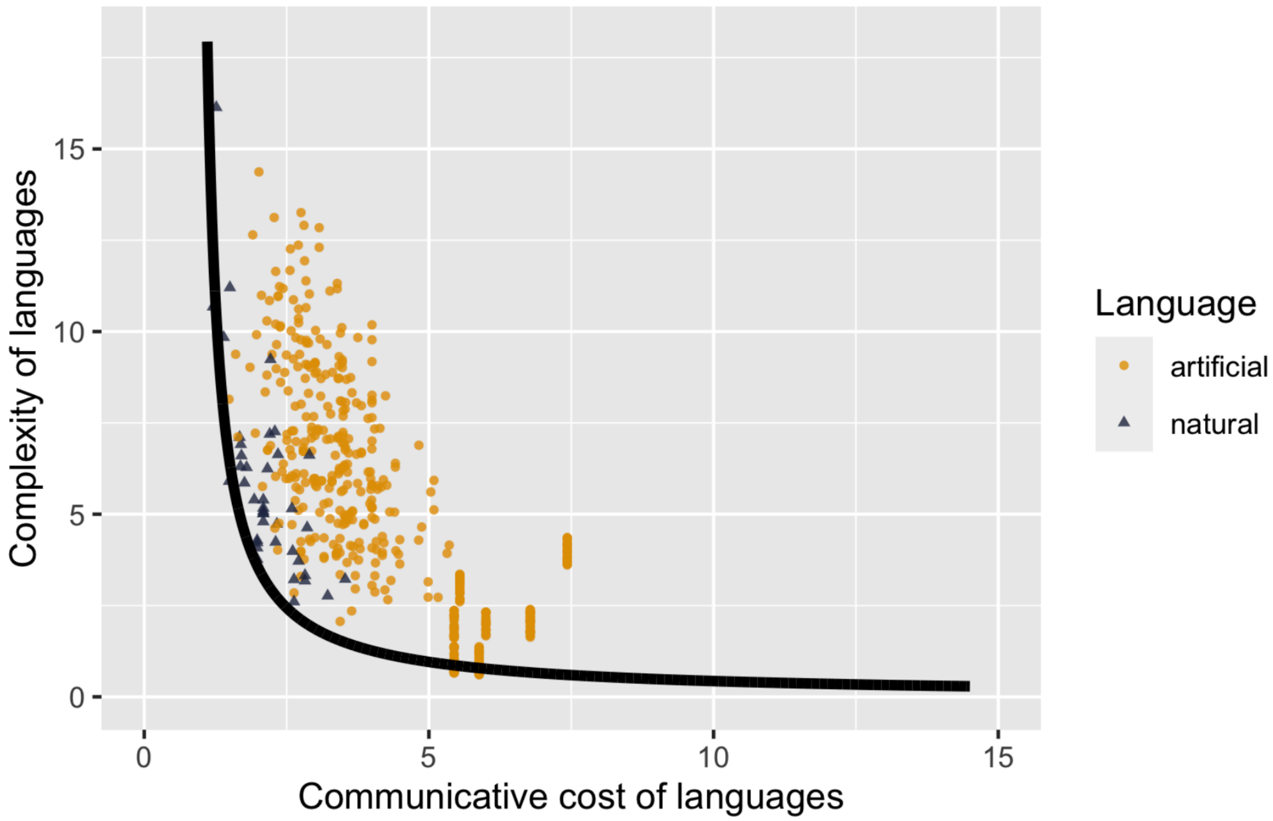
VIA SIMPLICITY

operator	type	gloss
$\cup$	SET $\times$ SET $\to$ SET	union
$\cap$	SET $\times$ SET $\to$ SET	intersection
$\setminus$	SET $\times$ SET $\to$ SET	setminus
$\iota(\cdot,\cdot)$	INT $\times$ SET $\to$ SINGLETON SET	'object at index'
$\lvert\cdot\rvert$	SET $\to$ INT	cardinality
$\subseteq$	SET $\times$ SET $\to$ BOOL	subset equal
$=$	INT $\times$ INT $\to$ BOOL	integer equality
$>$	INT $\times$ INT $\to$ BOOL	integer larger than
$\neg$	BOOL $\to$ BOOL	negation
$\wedge$	BOOL $\times$ BOOL $\to$ BOOL	and
$\vee$	BOOL $\times$ BOOL $\to$ BOOL	or

van de Pol, Lodder, van Maanen, Steinert-Threlkeld, Szymanik.
Quantifiers satisfying semantic universals have shorter minimal description length.
Cognition 2022

- Why do languages lexicalize only some possible concepts?
- Pick a LoT.
- Generate artificial concepts within the LoT.
- Lexicalized concepts have shorter MDL in the LoT than non-lexicalized, yet logically possible, ones.
- LLMs exhibit similar bias for simplicity (Wang et al., 2024)

SIMPLICITY VS INFORMATIVENESS



Indefinite pronoun	Flavors	Feature formula	$c(i)$
someone	specific known, specific unknown, non-specific	SE^-	1
anyone	negative polarity, free choice	$SE^+ \cap N^-$	2
no one	negative indefinite	N^+	1

Table 3: English indefinite pronouns, the flavors they convey, their feature formulae and their complexities.

Indefinite pronoun	Flavors	Feature formula	$c(i)$
kto-to	specific unknown, non-specific	$K^- \cap SE^-$	2
kto-nibud'	non-specific	$S^- \cap SE^-$	2
kto-libo	non-specific, negative polarity	$(S^- \cap SE^-) \cup ((SE^+ \cap R^+) \cap N^-)$	5
nikto	negative indefinite	N^+	1
koe-kto	specific known	K^+	1
kto by to ni bylo	negative polarity	$(SE^+ \cap R^+) \cap N^-$	3
kto ugodno	free choice	$SE^+ \cap R^-$	2

Table 4: Russian indefinite pronouns, the flavors they convey, their feature formulae and their complexities.

Denić, Steinert-Threlkeld, Szymanik. Indefinite pronouns optimize the simplicity/informativeness trade-off. Cognitive Science, 2022

- The complexity of the language system is the minimal number of rules needed to define it in a LoT.
- Communicative cost is the reconstruction error.
- Evolution balances complexity and the communicative costs.

BUT IS THERE LOT UNIFYING ALL THE EXAMPLES?

“

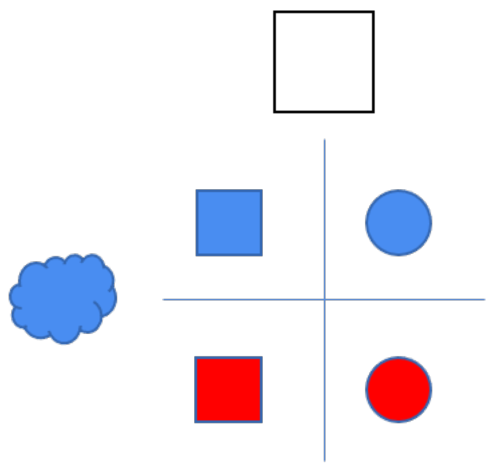
“The choice of innate primitives can be viewed as a strictly empirical question that should be determined through independent experiments.”.

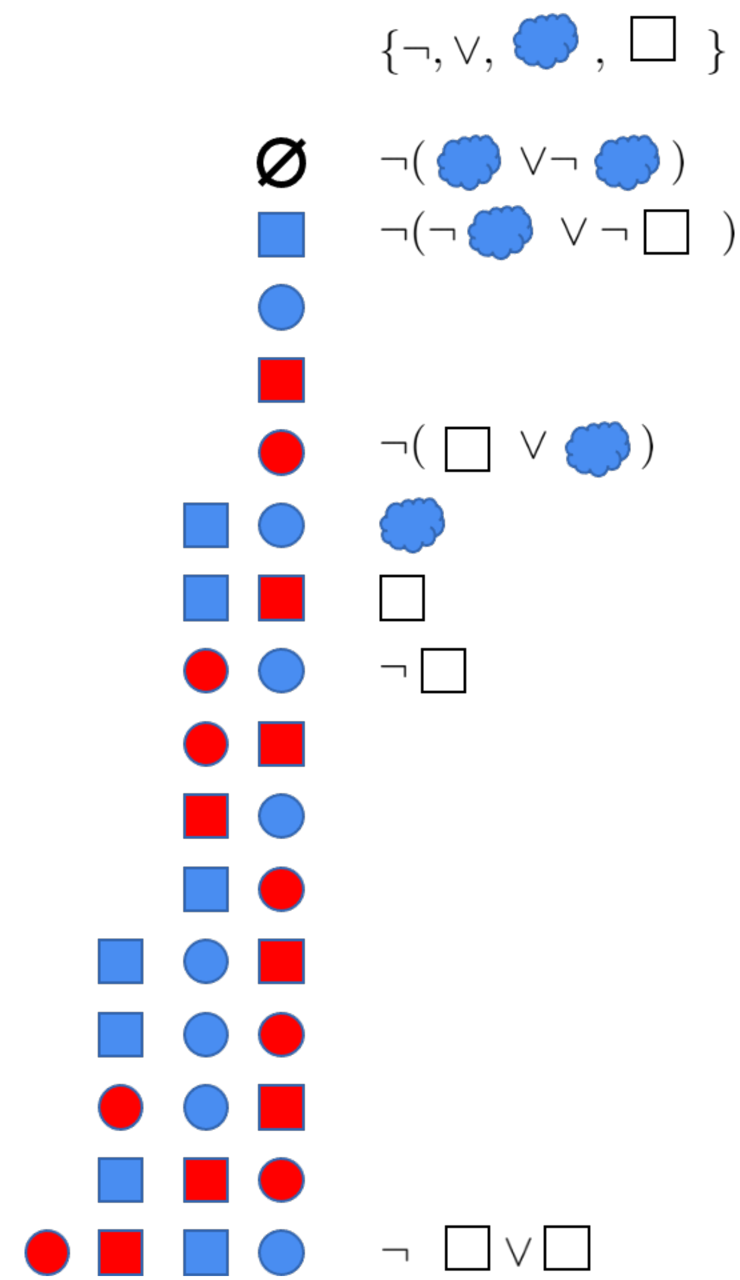
-Piantadosi & Jacobs, 2016

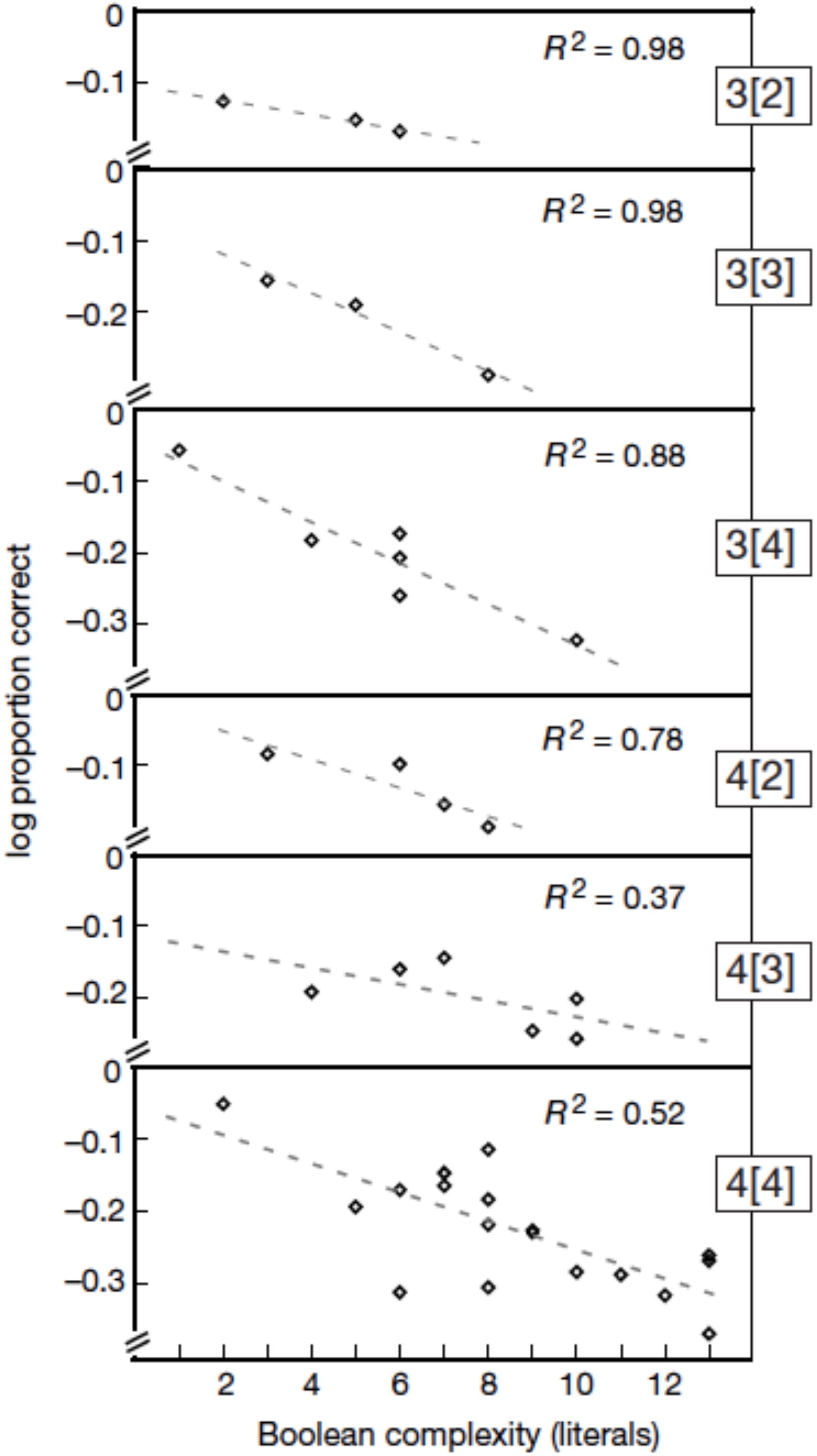
CAN WE INFER IT?

APPROACH 1: INFERRING LOT FROM LEARNING DATA

LEARNING A NEW CONCEPT, “GLEEB”







Feldman `01

FELDMAN'S RESULTS

- Consider an arbitrary Boolean concept defined by P positive examples over D binary features, P[D].
- Boolean complexity accounts for 50% of variance in the dataset.

QUESTIONS

- Which Boolean connectives?

WHAT'S THE RIGHT FELDMAN'S GRAMMAR?

SimpleBoolean		
START	→	*lambda x . BOOL*
BOOL	→	*(and BOOL BOOL)*
		(or BOOL BOOL)
		(not BOOL)
		true
		false
BOOL	→	*(F OBJECT)*
OBJECT	→	*x*
F	→	COLOR
		SHAPE
		SIZE
COLOR	→	*blue?*
		green?
		yellow?
SHAPE	→	*circle?*
		rectangle?
		triangle?
SIZE	→	*size1?*
		size2?
		size3?

NAND		
START	→	*lambda x . BOOL*
BOOL	→	*(nand BOOL BOOL)*
		true
		false
BOOL	→	*(F OBJECT)*
OBJECT	→	*x*
F	→	COLOR
		SHAPE
		SIZE
COLOR	→	*blue?*
		green?
		yellow?
SHAPE	→	*circle?*
		rectangle?
		triangle?
SIZE	→	*size1?*
		size2?
		size3?

WHAT’S THE RIGHT FELDMAN’S GRAMMAR?

DNF		
START	→	*lambda x . DISJ*
DISJ	→	*CONJ*
		(or CONJ DISJ)
CONJ	→	*BOOL*
		(and BOOL CONJ)
BOOL	→	*(F OBJECT)*
		(not (F OBJECT))
OBJECT	→	*x*
F	→	COLOR
		SHAPE
		SIZE
COLOR	→	*blue?*
		green?
		yellow?
SHAPE	→	*circle?*
		rectangle?
		triangle?
SIZE	→	*size1?*
		size2?
		size3?

HORN CLAUSE		
START	→	*lambda x . HORN-CONJ*
HORN-CONJ	→	*HORN-CLAUSE*
		(and HORN-CLAUSE HORN-CONJ)
HORN-CLAUSE	→	*(implies HORN-CONJ PRIM)*
HORN-CLAUSE	→	*(implies HORN-CONJ false)*
PRIM	→	*(F OBJECT)*
OBJECT	→	*x*
F	→	COLOR
		SHAPE
		SIZE
COLOR	→	*blue?*
		green?
		yellow?
SHAPE	→	*circle?*
		rectangle?
		triangle?
SIZE	→	*size1?*
		size2?
		size3?

GRAMMAR COMPARISON

- Bayesian data analysis model: which representational system is the most likely, given human responses?

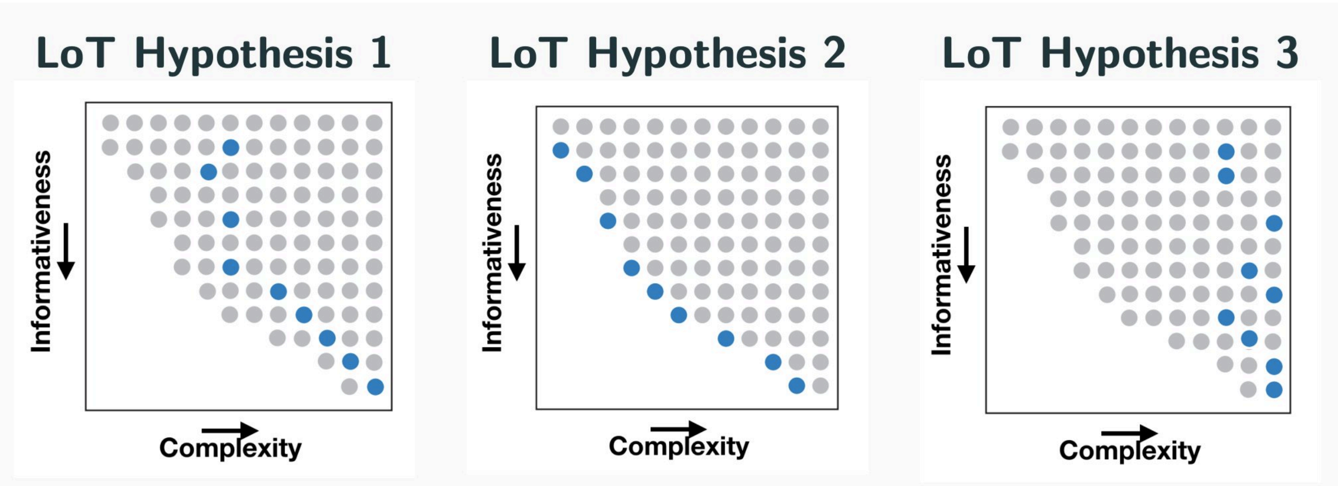
Grammar	H.O.LL	FP	$R^2_{response}$	R^2_{mean}
FullBoolean	−16296.84	27	.88	.60
Biconditional	−16305.13	26	.88	.64
CNF	−16332.39	26	.89	.69
DNF	−16343.87	26	.89	.66
SimpleBoolean	−16426.91	25	.87	.70
Implies	−16441.29	26	.87	.70
HornClause	−16481.90	27	.87	.65
NAND	−16815.60	24	.84	.61
NOR	−16859.75	24	.85	.58
uniform	−19121.65	4	.77	.06
exemplar	−23634.46	5	.55	.15
OnlyFeatures	−31670.71	19	.54	.14
Response-Biased	−37912.52	4	.03	.04

Piantadosi et al. `16

APPROACH 2: INFERRING LOT-PRIMITIVES FROM THE COMPLEXITY-INFORMATIVENESS TRADE-OFF

FROM TRADE-OFF

➤ What LoT primitives underlie number concepts 1-99?



Denoted number (numeral)	Morphosyntactic make-up
6 (*iwan*)	10 (*-wan*) − (∅) 4 (*i-*)
30 (*wanetuhotne*)	2 (*-tu-*) · (∅) 20 (*-hotne*) − (*-e-*) 10 (*wan-*)
42 (*tuikashimatuhotne*)	2 (*-tu-*) · (∅) 20 (*-hotne*) + (*-ikashima-*) 2 (*tu-*)

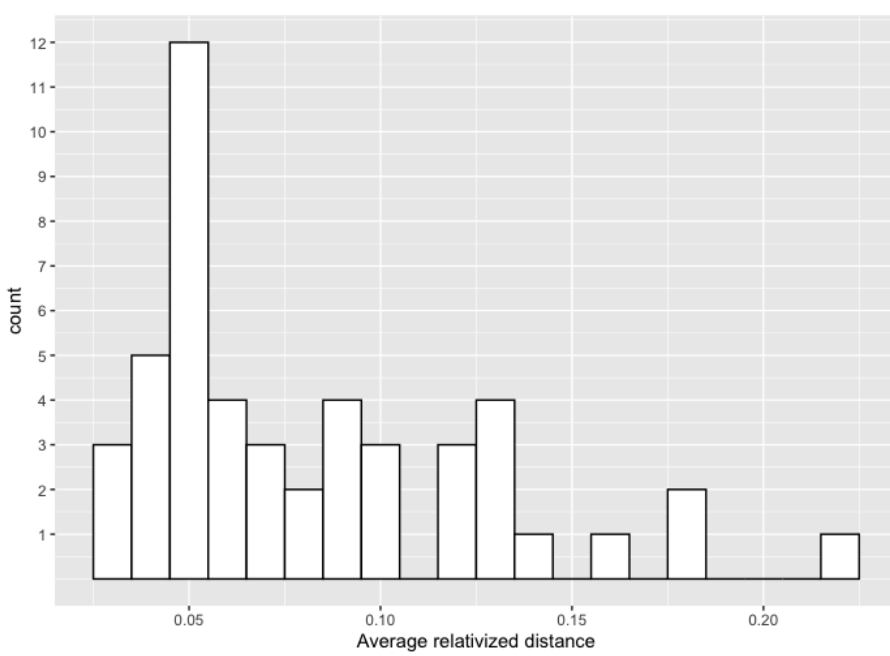


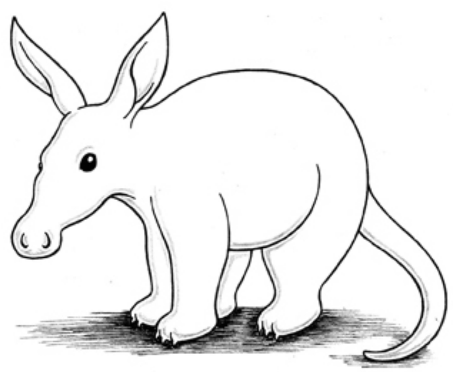
Table 2: Top three LoT hypotheses

PRIM
{1, 2, 3, 5, 10}
{1, 2, 3, 4, 5, 10}
{1, 2, 5, 10}

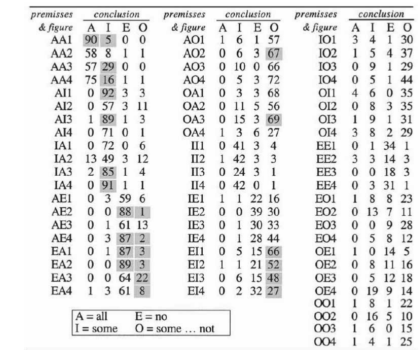
Denić, Szymanik. Reverse-engineering the language of thought: A new approach. CogSci, 2022
See also: Denić, Szymanik. Recursive Numeral Systems Optimize the Trade-off Between Lexicon Size and Average Morphosyntactic Complexity. Cognitive Science, 2024.

APPROACH 3: INFERRING LOT-RULES FROM REASONING DATA

Data - syllogistic Reasoning



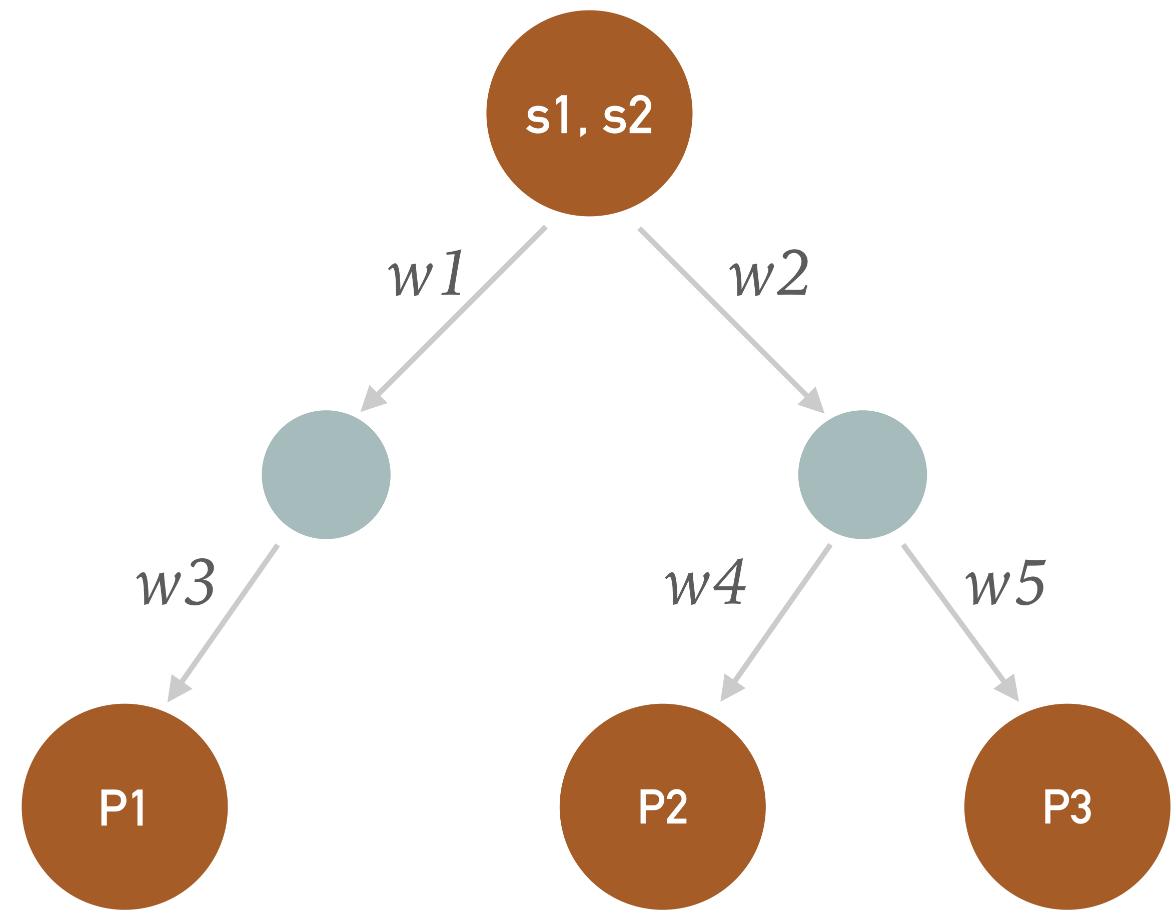
1. *All aardvarks are insectivores.*
2. *All Orycteropodidae are aardvarks.*
3. ***90%:*** *All Orycteropodidae are insectivores.*
4. ***5%:*** *Some Orycteropodidae are insectivores.*
5. ***5%:*** *Others, including erroneous.*



premisses & figure	*conclusion* A	I	E	O	*premisses & figure*	*conclusion* A	I	E	O	*premisses & figure*	*conclusion* A	I	E	O
AA1	90	5	0	0	AO1	1	6	1	57	IO1	3	4	1	30
AA2	58	8	1	1	AO2	0	6	3	67	IO2	1	5	4	37
AA3	57	29	0	0	AO3	0	10	0	66	IO3	0	9	1	29
AA4	75	16	1	1	AO4	0	5	3	72	IO4	0	5	1	44
AI1	0	92	3	3	OA1	0	3	3	68	OI1	4	6	0	35
AI2	0	57	3	11	OA2	0	11	5	56	OI2	0	8	3	35
AI3	1	89	1	3	OA3	0	15	3	69	OI3	1	9	1	31
AI4	0	71	0	1	OA4	1	3	6	27	OI4	3	8	2	29
IA1	0	72	0	6	II1	0	41	3	4	EE1	0	1	34	1
IA2	13	49	3	12	II2	1	42	3	3	EE2	3	3	14	3
IA3	2	85	1	4	II3	0	24	3	1	EE3	0	0	18	3
IA4	0	91	1	1	II4	0	42	0	1	EE4	0	3	31	1
AE1	0	3	59	6	IE1	1	1	22	16	EO1	1	8	8	23
AE2	0	0	88	1	IE2	0	0	39	30	EO2	0	13	7	11
AE3	0	1	61	13	IE3	0	1	30	33	EO3	0	0	9	28
AE4	0	3	87	2	IE4	0	1	28	44	EO4	0	5	8	12
EA1	0	1	87	3	EI1	0	5	15	66	OE1	1	0	14	5
EA2	0	0	89	3	EI2	1	1	21	52	OE2	0	8	11	16
EA3	0	0	64	22	EI3	0	6	15	48	OE3	0	5	12	18
EA4	1	3	61	8	EI4	0	2	32	27	OE4	0	19	9	14
										OO1	1	8	1	22
										OO2	0	16	5	10
										OO3	1	6	0	15
										OO4	1	4	1	25

A = all E = no
I = some O = some … not

- All-Some: `All A are B' implies `Some A are B'.
- No-Some not: `No A are B' implies `Some A are not B'.
- Conversion1: `Some A are B' implies `Some B are A';
- Conversion2: `No A are B' implies `No B are A".
- Monotonicity rule.



Zhai, Titov, Szymanik. Toward a probabilistic mental logic for the syllogistic fragment of natural language. Amsterdam Colloquium, 2015

HOW MUCH SUCCESS SHOULD WE EXPECT IN INFERRING PRIMITIVES?

STRATEGY

- Pick a conceptual domain.
- Define a space of possible LoTs.
- Define a way that LoT can influence behaviour, e.g., through category learning via a simplicity bias.
- Run simulated experiments with known LoTs and see whether it's possible to recover the underlying LoT from the simulated behavioural data accurately.

MODEL

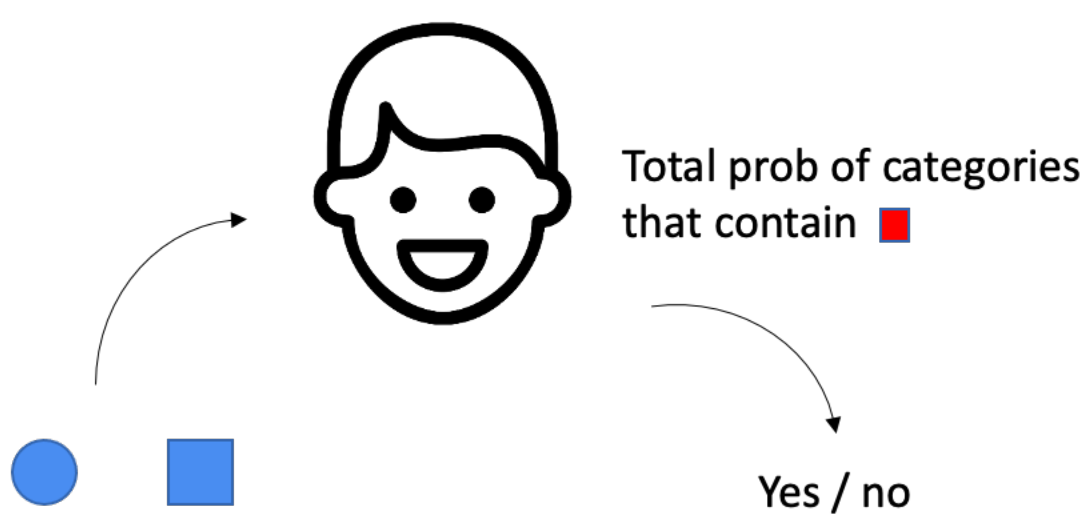
Table 1
Glossary of technical terms and symbols as used in the computational model.

Property:	A binary property, e.g. 'being red'.
Object:	A set of properties.
Category:	A set of objects.
Operator:	A function from Booleans to a Boolean.
Language of Thought (LoT):	A set of operators.
p, q, r, s:	The four properties in the model.
$\mathcal{O}$:	The ordered tuple of 16 objects.
Ω:	The set of 9 operators.

- 4 binary properties
- An object is a set of properties.
- A category is a set of objects
- LoTs are functionally complete, non-redundant subsets of 16 Boolean binary operators.

CATEGORY GENERALIZATION DESIGN

$$p(\text{category} \mid \bullet\,\blacksquare\,) \propto \overset{\text{Strong sampling}}{p(\,\bullet\,\blacksquare \mid \text{category})}\,\overset{\text{Simplicity prior}}{p(\text{category})}$$



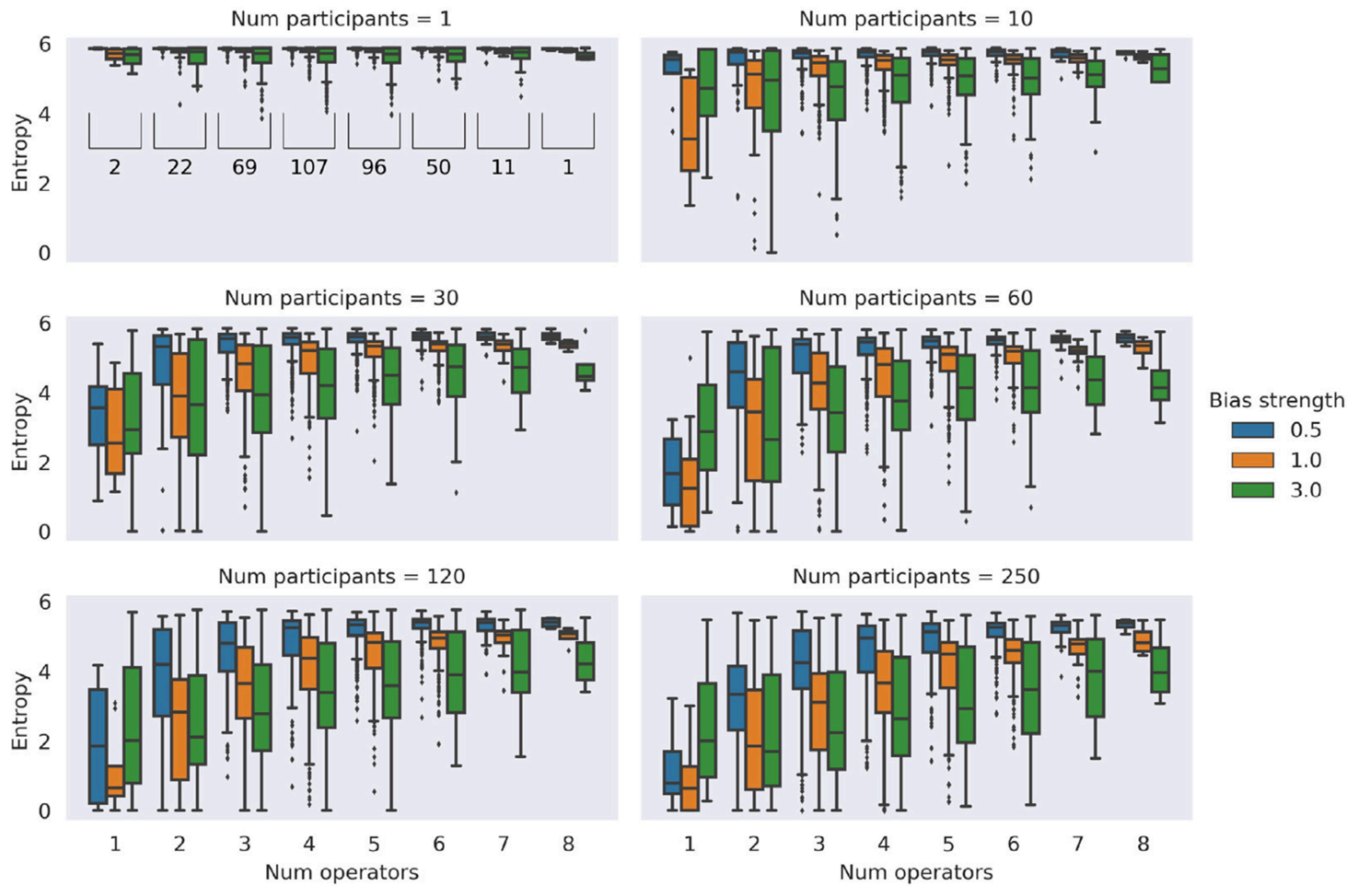
SIMULATED EXPERIMENT 1

- The agent sees some examples of objects in the ‘true’ category (which is, in fact, a random selection of possible objects).
- The agent calculates the posterior over categories given the observed objects and the agent’s true LoT.
- Then the agent needs to decide whether the remaining objects belong to the category (by summing the posterior probability of all categories that contain the object).
- As experimenters, we calculate the simulated experimenter’s posterior over LoT given the participants’ categorization data.
- Prior is given by the minimal formula in the LoT for a category.

25776 SIMULATED EXPERIMENTS

Name	Abbreviation	Symbol
Conjunction (and)	A	$\wedge$
Disjunction (or)	O	$\vee$
Conditional	C	$\rightarrow$
Negated conditional	NC	$\nrightarrow$
Biconditional	B	$\leftrightarrow$
Negated biconditional	XOR	$\nleftrightarrow$
Negated conjunction	NAND	$\barwedge$
Negated disjunction	NOR	$\veebar$
Negation	N	$\neg$

- 4 properties
- 358 possible LoTs
- 65536 categories
- Number of examples given in the experiment: 1, 5, 10, 15
- Number of participants: 1, 10, 30, 60, 120, 250
- Simplicity bias strength: 0.5, 1, 3
- What we look at for each combination of parameter values is the distribution of the simulated experimenter’s posterior entropies across LoTs.



LoT	# recovered	LoT	# recovered
$\wedge, \neg$	19	$\wedge, \nleftrightarrow, \not\wedge$	3
$\vee, \neg$	14	$\leftrightarrow, \not\vee$	3
$\wedge, \not\vee$	14	$\vee, \neg, \nrightarrow$	3
$\not\vee$	11	$\neg, \rightarrow$	3
$\wedge, \not\vee, \nrightarrow$	8	$\vee, \wedge, \neg, \nrightarrow$	3
$\neg, \not\vee$	6	$\wedge, \neg, \nleftrightarrow$	3
$\wedge, \neg, \leftrightarrow$	5	$\wedge, \neg, \not\vee$	3
$\vee, \neg, \leftrightarrow$	5	$\not\wedge, \nrightarrow$	2
$\vee, \not\vee$	5	$\vee, \wedge, \not\vee$	2
$\not\wedge$	4	$\wedge, \leftrightarrow, \not\vee, \nrightarrow$	2
$\leftrightarrow, \nrightarrow$	4	$\neg, \nleftrightarrow, \nrightarrow$	2
$\neg, \nrightarrow$	4	$\nleftrightarrow, \not\wedge, \not\vee$	2
$\vee, \wedge, \neg$	4	$\wedge, \leftrightarrow, \not\vee$	2
$\vee, \not\vee, \nrightarrow$	3	$\wedge, \neg, \rightarrow$	2
$\not\vee, \nrightarrow$	3	$\vee, \neg, \nleftrightarrow$	2

Table 4: 30 LoTs that were recovered most often, along with the number of times they were recovered across all simulated experiments. This gives a sense of what LoTs are easiest to recover amongst all LoTs.

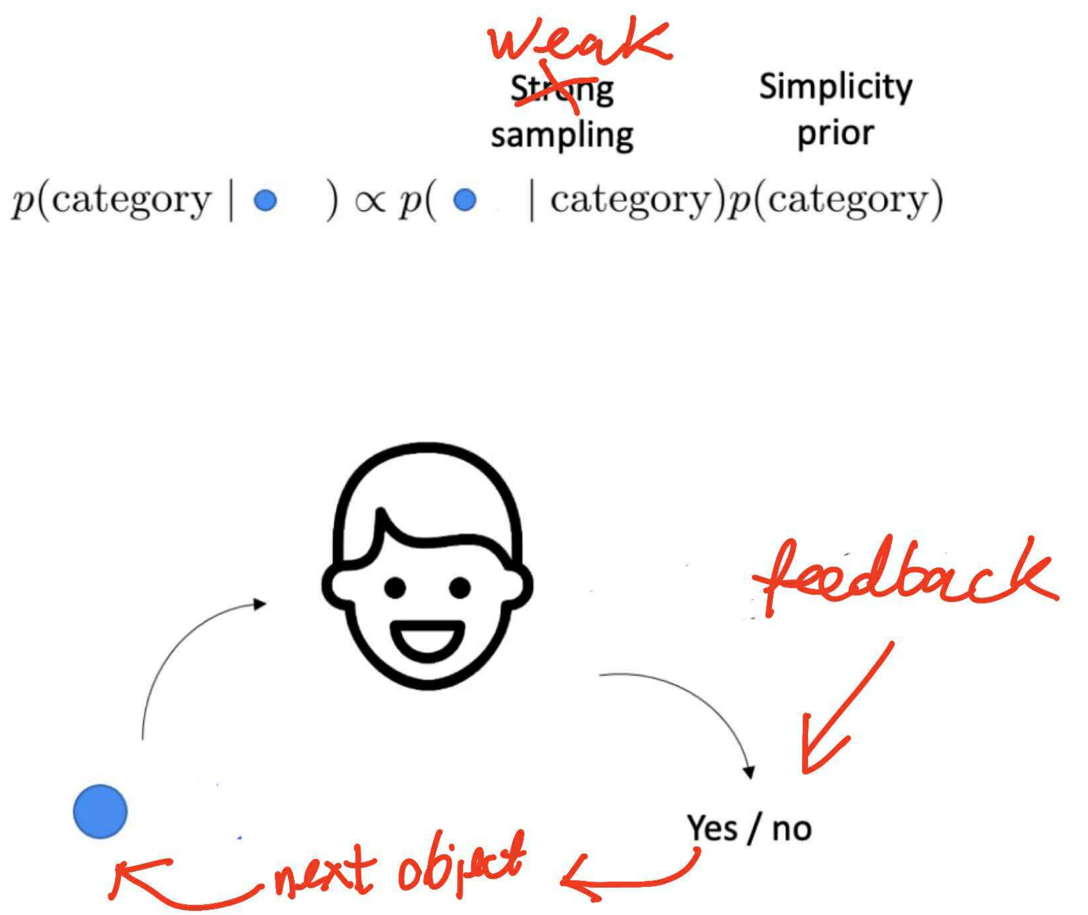
RESULTS

- The more participants, the better the recovery.
- The stronger the simplicity bias, the more recoverable LoTs are in general.
- The fewer operators in the LoT, the more recoverable it is.
- LoTH may be empirically productive only if the postulated LoTs are simple.
- Therefore, we require more and better theory, not just brute force search within the LoT space.

Carcassi & Szymanik. The Boolean Language of Thought is recoverable from learning data, Cognition, 2023.

SERIAL AND DYNAMIC DESIGN

tracking learning trajectory



SIMULATED EXPERIMENT 2

- The agent sees each of 16 objects, judges whether it belongs to the category, and gets feedback.
- So, agents receive both positive and negative evidence.
- Serial design: we select both category and order by randomly sampling
- Dynamic design: we choose the next object to show using the greedy Bayesian Optimal Design
- Optional stopping: up to 250 participants or posterior probability > 0.95

1432 SIMULATED EXPERIMENTS

- 4 properties
- 358 possible LoTs
- 65536 categories
- Simplicity bias strength: 0.5, 1, 3
- Serial and dynamic design

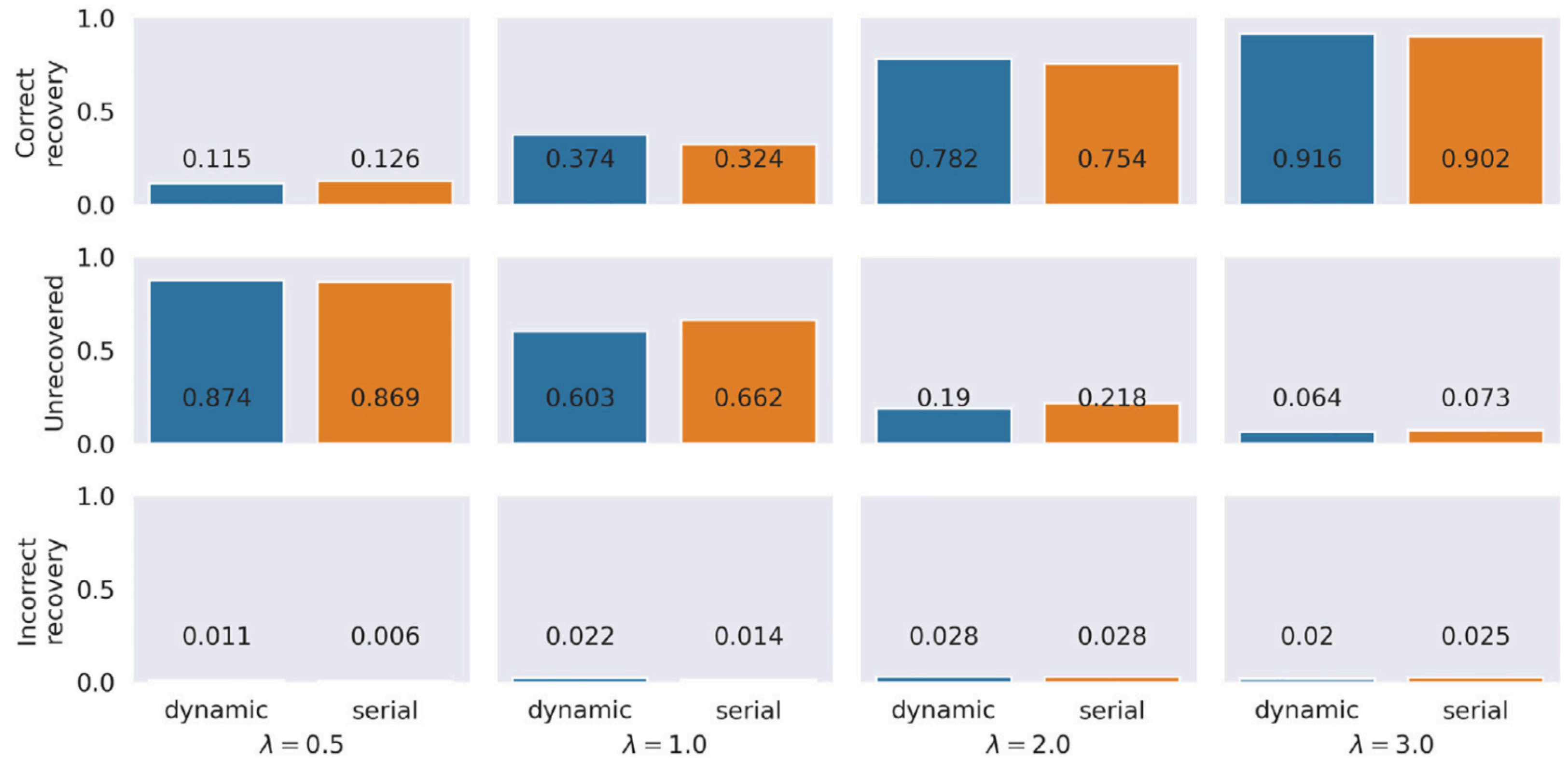


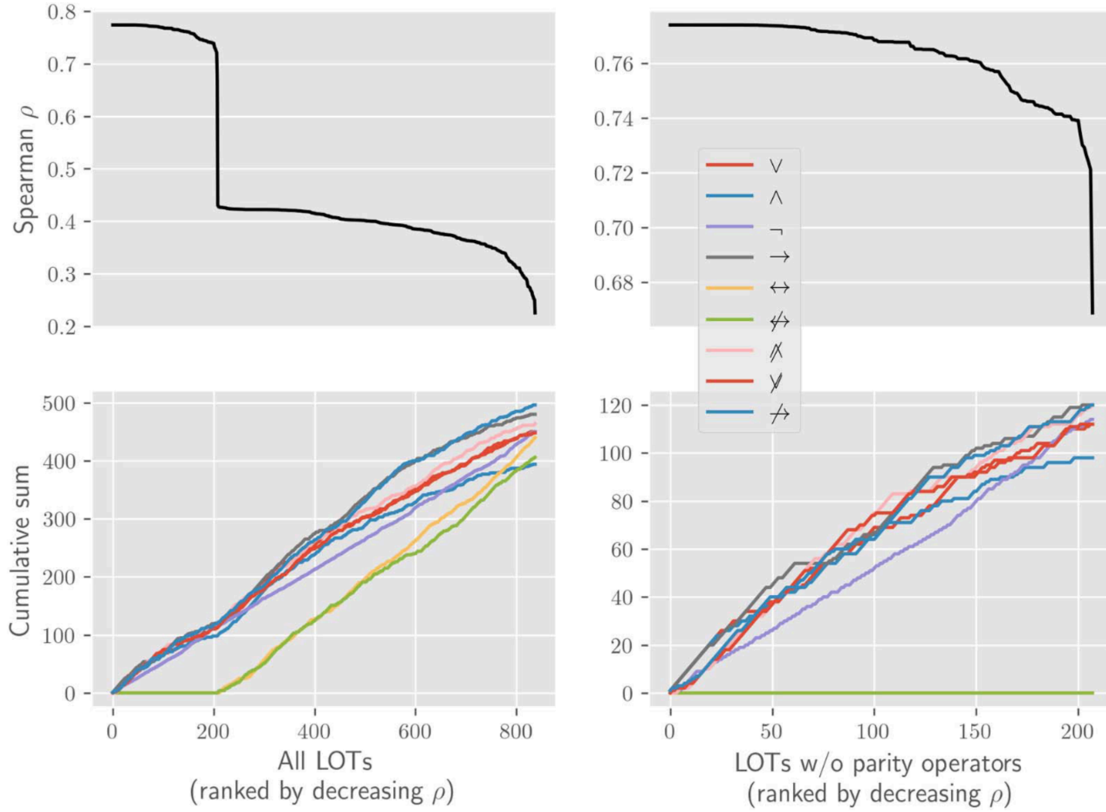
Fig. 7. Proportions of each outcome out of the 358 experiments that were ran for each level of simplicity bias strength λ. Results are shown both for the dynamic and for the serial design. The main result is that with these more sophisticated designs, the true LoT can reliably be recovered from learning data, as long as the simplicity bias is strong enough.

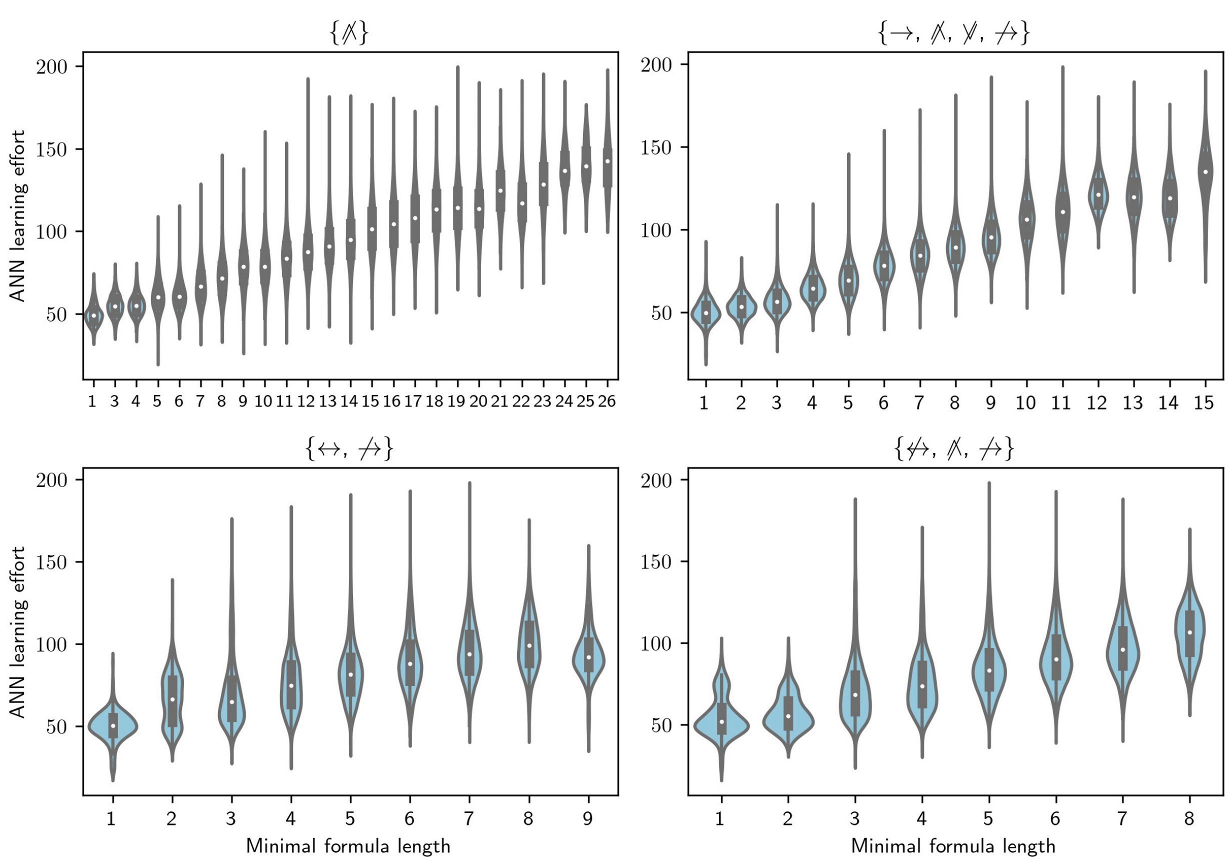
RESULTS

- With a strong simplicity bias, the majority of the true LoTs are recovered.
- The misidentification is low.
- Dynamic design doesn't help.
- The fewer operators in the LoT, the more recoverable it is.
- LoTH may be a productive endeavor, but doing it in reality will be much harder

POST-SCRIPTUM: WHAT IF COGNITION IS ALL NON-SYMBOLIC?

DOES THAT EVEN MATTER?





Carcassi & Szymanik. Neural Networks track the logical complexity of Boolean concepts, Open Mind 2022.

LOT VS ANNS

- Do LoTH and connectionism have the same empirical import in the domain of categorization?
- Do they make the same predictions about the effort required to acquire categories?
- LoT: The categories with the shortest minimal formulas are the easiest.
- Connectionism: the average loss across epochs and batches
- There is an overall positive rank correlation between logical complexity and ANN learning effort

CONCLUSIONS

- LoT is often the engine of computational cognitive models
- Via the notions of complexity/simplicity prior
- Leading to interesting theoretical and empirical insights
- To unify those models, we need to recover the true LoT
- Such LoT could help AI get closer to human-like intelligence

THANK YOU!