Toward Probabilistic Natural Logic for Syllogistic Reasoning

Fangzhou Zhai, Jakub Szymanik and Ivan Titov

Institute for Logic, Language and Computation, University of Amsterdam

Abstract

Natural language contains an abundance of reasoning patterns. Historically, there have been many attempts to capture their rational usage in normative systems of logical rules. However, empirical studies have repeatedly shown that human inference differs from what is characterized by logical validity. In order to better characterize the patterns of human reasoning, psychologists have proposed a number of theories of reasoning. In this paper, we combine logical and psychological perspectives on human reasoning. We develop a framework integrating Natural Logic and Mental Logic traditions. We model inference as a stochastic process where the reasoner arrives at a conclusion following a sequence of applications of inference steps (both logical rules and heuristic guesses). We estimate our model (i.e. assign weights to all possible inference rules) on a dataset of human syllogistic inference while treating the derivations as latent variables in our model. The computational model is accurate in predicting human conclusions on unseen test data (95% correct predictions) and outperforms other previous theories. We further discuss the psychological plausibility of the model and the possibilities of extending the model to cover larger fragments of natural language.

1 Introduction

1.1 Syllogistic Reasoning

The psychology of reasoning tries to answer one fundamental question: how do people reason? This question is also central for many other scientific disciplines, from linguistics and economics to cognitive science and artificial intelligence[1]. Logic was first to study reasoning systematically and Aristotle proposed the syllogistic theory as an attempt to normatively characterize rationality. And even though modern logic has developed intricate theories of many fragment of natural language, the syllogistic fragment continuously receives attention from researchers (see [11] for a review of the theories of syllogisms). The sentences of syllogisms are of four different sentence types (or 'moods'), namely:

$$\text{All } A \text{ are } B : \text{universal affirmative (A)}$$
$$\text{Some } A \text{ are } B : \text{particular affirmative (I)}$$
$$\text{No } A \text{ are } B : \text{universal negative (E)}$$
$$\text{Some } A \text{ are not } B : \text{particular negative(O)}$$

Each syllogism has two sentences as the premises, and one as the conclusion. Traditionally, according to the arrangements of the terms in the premises, syllogisms are classified in to four categories, or 'figures':

---

[1]See, e.g., [8] for a survey of logic and cognitive science.

	Figure 1	Figure 2	Figure 3	Figure 4
	BC	CB	BC	CB
	AB	AB	BA	BA
	AC	AC	AC	AC

Syllogisms are customarily identified by their sentence types and figures. For example, 'AI3E' refers to the syllogism whose premises are of sentence types A and I, and whose terms are arranged according to figure 3, and whose conclusion is of type E. Therefore, altogether, 'AI3E' refers to the following syllogism:

$$\text{All } B \text{ are } C$$
$$\text{Some } B \text{ are } A$$
$$\overline{\qquad\qquad\qquad}$$
$$\text{No } A \text{ are } C$$

As there are four different sentence types and four different figures, there are 256 equivalent syllogisms in total. These syllogisms are also referred to as the ones that follows the scholastic order. Of those 24 are valid according to the semantics of traditional syllogistic logic, and 15 of these 24 are valid according to the semantics of modern predicate logic.

Psychologists have developed a battery of experimental tests to study human syllogistic reasoning. In one typical experimental design, the reasoners are presented with the premises and asked 'What follows necessarily from the premises?'. Chater and Oaksford [3] have compared five experimental studies of this sort and computed the weighted average of the data, that is the percentage that each conclusion was drawn. The data is shown in Table 1.

One important observation made by Chater and Oaksford [3] is that logical validity seems to be a crucial factor for an explanation of the participants' performance. Firstly, the average percentage of reasoners arriving at a valid conclusion is 51%, while that of arriving at an invalid conclusion is 11%: it seems that participants indeed made an effort along the path of validity. Secondly, reasoners tends to mistakenly arrive at invalid syllogisms that are different from valid ones just by their figures. For example, the AO2O syllogism is the only valid one among the four AOO syllogisms, however, reasoners endorse the other three AOO syllogisms (namely AO1O, AO3O and AO4O) with fairly high probability. This might be a sign that people are actually not that bad at logic (see, e.g., [6]): even if an error is made, the most probable wrongly endorsed syllogism is quite similar to a valid one, which differs only in the figure. Thirdly, the mean entropy of the syllogistic premises that yields at least one valid conclusion, according to the table above, is 0.729, however, that of the ones that yield no valid syllogisms is 0.921. The difference indicates that the psychological procedures triggered by the two groups of premises are likely to be different.

1.2   Mental Logic

Rips [13] has proposed a theory of quantified reasoning based on formal inference rules. The underlying psychological assumption is that logical formulas can be used as the mental representations of reasoning steps and that the inference rules are the basic reasoning operations of the mind. Rips has argued that, deductive reasoning, as a psychological procedure, is the generation of 'a set of sentences linking the premises to the conclusion', and 'each link is the embodiment of an inference rule that reasoners consider intuitively sound'. He has formulated a set of rules that includes both sentential connectives and quantifiers and implemented such system as a computational mode PSYCOP.

Syllogism	Conclusion					Syllogism	Conclusion				
	A	I	E	O	NVC		A	I	E	O	NVC
AA1	**90**	**5**	0	0	5	AO1	1	6	1	57	35
AA2	58	8	1	1	32	AO2	0	6	3	**67**	24
AA3	57	**29**	0	0	14	AO3	0	10	0	66	24
AA4	75	**16**	1	1	7	AO4	0	5	3	72	20
AI1	0	92	3	3	2	OA1	0	3	3	68	26
AI2	0	57	3	11	29	OA2	0	11	5	56	28
AI3	1	**89**	1	3	7	OA3	0	15	3	**69**	13
AI4	0	71	0	1	28	OA4	1	3	6	27	63
IA1	0	72	0	6	22	II1	0	41	3	4	52
IA2	13	49	3	12	23	II2	1	42	3	3	51
IA3	2	**85**	1	4	8	II3	0	24	3	1	72
IA4	0	**91**	1	1	7	II4	0	42	0	1	57
AE1	0	3	59	6	32	IE1	1	1	22	16	60
AE2	0	0	88	**1**	11	IE2	0	0	39	30	31
AE3	0	1	61	13	25	IE3	0	1	30	33	36
AE4	0	3	**87**	**2**	8	IE4	0	42	0	1	57
EA1	0	1	**87**	**3**	9	EI1	0	5	15	**66**	14
EA2	0	0	**89**	**3**	8	EI2	1	1	21	**52**	25
EA3	0	0	64	**22**	14	EI3	0	6	15	**48**	31
EA4	1	3	61	**8**	28	EI4	0	2	32	**27**	39
OE1	1	0	14	5	80	OO1	1	8	1	12	78
OE2	0	8	11	16	65	OO2	0	16	5	10	69
OE3	0	5	12	18	65	OO3	1	6	0	15	78
OE4	0	19	9	14	58	OO4	1	4	1	25	69
IO1	3	4	1	30	62	OI1	4	6	0	35	55
IO2	1	5	4	37	53	OI2	0	8	3	35	54
IO3	0	9	1	29	61	OI3	1	9	1	31	58
IO4	0	5	1	44	50	OI4	3	8	2	29	58
EE1	0	1	34	1	64	EO1	1	8	8	23	60
EE2	3	3	14	3	77	EO2	0	13	7	11	69
EE3	0	0	18	3	78	EO3	0	0	9	28	63
EE4	0	3	31	1	65	EO4	0	5	8	12	75

Table 1: Percentage of times each syllogistic conclusions was endorsed. The data is from a meta-analysis in [3]. 'NVC' stands for 'No Valid Conclusion', all numbers have been rounded to the closest integer. A bold number indicates that the corresponding conclusion is logically valid.

The reasoner modeled by the theory derives only but not all logically valid conclusions (i.e., it is logically sound but not complete). It puts constraints on the application of the inference rules to deal away with logical omniscience: certain logical truths are not derivable in Mental Logic theory. Instead of accepting standard proof-theoretical system, Rips has selected the inference rules that seem psychologically 'primitive', even if derivable from other rules. Nevertheless, the model still uses arbitrary abstract rules and formal representations (roughly corresponding to the natural deduction system for first-order logic). Moreover, the model, by its mere design, cannot explain reasoning mistakes (see also [9]).

1.3 Natural Logic

Due to psychological, computational and linguistic influences, some of the normative inference rules have been adapted to natural language as a part of so-called Natural Logic Program [15, 2]. Contrary to the Mental Logic of Rips, the Natural Logics identify valid inferences by their lexical and syntactic features, without requiring a full semantic interpretation. For example, some natural language quantifiers are upward monotone in their first argument, like the quantifier 'some'. It means that the inference from 'Some pines are green' to 'Some plants are green' is valid since all pines are plants. The 'pines' can be actually replaced by any object that contains all pines. People can reason based on monotonicity even when the underlying meaning of terms is unclear for them. For example, from 'Every Dachong has nine beautiful tails' people would infer 'Every Dachong has nine tails', without knowing the meaning of 'Dachong' (which simple means 'tiger' in Chinese). In a way, monotonicity operates on the surface of natural language.

Using ideas from Natural Logic, Geurts [6] has designed a proof system for syllogistic reasoning that pivots on the notion of monotonicity. Geurts' proof system for syllogistic reasoning consists of the following set of rules R:

· *All-Some*: 'All A are B' implies 'Some A are B'.

· *No-Some_not*: 'No A are B' implies 'Some A are not B'.

· *Conversion*: 'Some A are B' implies 'Some B are A'; 'No A are B' implies 'No B are A'.

· *Monotonicity*: If A entails B, then the A in any upward entailing position can be substituted by a B, and the B in any downward entailing position can be substituted by an A.

Geurts has further enriched the proof system with difficulty weights assigned to each inference rules to evaluate the difficulty of valid syllogistic reasoning. Geurts assumed that different rules cost different amount of cognitive resources. He gives each reasoner an initial budget of 100 units; each use of the monotonicity rule costs 20 units; a proof containing a 'Some Not' proposition costs an additional 10 units. Taking the remaining budget as an evaluation of the difficulty of each syllogism, the evaluation system fits the experimental data from [3] well. However, the system cannot make any evaluation on most invalid syllogisms, hence cannot explain why reasoners can possibly arrive at invalid conclusions.[2]

2 Data-driven Probabilistic Natural Logic for Syllogistics

2.1 Approach

In this paper we design and estimate a computational model for syllogistic reasoning based on a probabilistic natural logic.[3] This can be treated as a first step to integrate the Mental Logic approach and the Natural Logic approach. It improves upon Mental Logic approach by substituting formal abstract inference rules with Natural Logic operating on the surface structure of Natural Language. That means, the mental representations are given directly as

---

[2]This is not a criticism of [6]. According to Geurts, the system was never intended to give a 'full-blown account of syllogistic reasoning' in the first place, see also [11].

[3]Compare with [5], where the authors designed probabilistic semantic automata for quantifiers whose parameters are also determined by the experimental data.

$$\boxed{\text{All } C \text{ are } B; \text{No } B \text{ are } A}$$

All−Some   *Terminate*

$$\boxed{\text{All } C \text{ are } B; \text{No } B \text{ are } A; \textbf{Some C are B}}$$ $$\boxed{\text{All } C \text{ are } B; \text{No } B \text{ are } A; \textbf{Nothing Follows}}$$
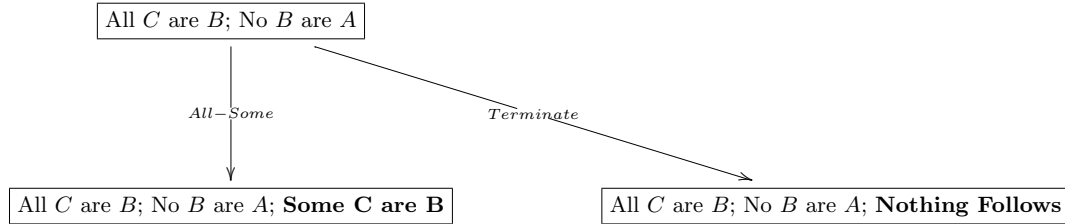
Figure 1: The Mental Representations

natural language sentences, without an intermediate layer of an abstract formal language. Our starting point is the logic developed by Geurts in [6] (see Section 1.3).

We assume that the procedure of reasoning consists of two types of mental events: the inferences made by the reasoners, which are deliberate and precise, and the guesses, which could be less reliable but fast. Accordingly, the model consists of two parts: the inference part, which takes the form of a probabilistic natural logic (i.e., the inference rules are weighted with probabilities) and the guessing part, which leads the reasoner to a possible conclusion in one step depending on a few heuristics. We implemented the model, and estimated it on the experimental data. The model is accurate at predicting human conclusions on unseen syllogisms (including mistakes) and the results yield interesting psychological implications.

2.2 Mental Representation

Similar to Rips' [13] proposal, we take the set of syllogistic sentences as the mental representation of reasoning. Namely, the reasoner maintains a set of sentences in the working memory to represent the state of reasoning, or more specifically, the reasoner keeps a record of the sentences that he considers true at the moment. We will refer to each representation as a **state**. Reasoning operations change the mental states. When performing reasoning, the reasoner generates a sequence of states in the working memory, where the initial state is the set of premises, and the final state contains the conclusion. These states are linked by the **reasoning events**, which can be a specific adoption of an inference rule. For example, given the 'AE4' premises, if the reasoner adopt the 'All - Some' rule (i.e., 'All A are B' implies 'Some A are B') on the premise 'All C are B', a 'Some C are B' will be obtained, possibly as a conclusion. The reasoner may also terminate the reasoning and decide that 'nothing follows', see Figure 1.

We would like to point out here that mental states may not be logically consistent. There are many reasons for this assumption. For example, people tend to adopt illicit conversions which often lead to the inconsistency. After all, people do often make mistakes resulting in conclusions that are inconsistent with assumptions, even while reasoning in a conscious, deliberate way (see, e.g., [10]).

2.3 Statistical Model of Reasoning Procedure

We formulate a generative probabilistic model of reasoning. First, reasoners conduct formal inferences, adopting possible logical rules with different probabilities (related to the cognitive

difficulty of the rule or some sort of reasoning preference). Each inference rule, $r \in R$ is adopted with a different probability specified by the associated weight w_r (*a tendency parameter*) which is estimated from the data. Formally, a probability of transitioning from state S to state S_r using a specific application of rule r is given by:

$$p(r|S, \mathbf{w}) = \frac{w_r}{w_G + \sum_{r' \in R} c_{r'} \cdot w_{r'}},$$

where $c_{r'}$ is the number of different possibilities how the rule r' can be adopted in the given state S and $\mathbf{w}$ is the vector of all tendency parameters. The parameter w_G reserves probability mass for 'terminating' the inference at state S and making a heuristic guess.

The reason to turn to the guessing scenario may have to do with the complexity of inference or the reasoner doubting the conclusion that was already obtained. When the reasoner enters the guessing scenario, the probability that the reasoner guesses 'nothing follows' is negatively correlated with the **informativeness** level (see [4]) of the premises, i.e., the amount of information that the premises carries: the more informative the premise, the less faith the reasoner have for a 'nothing follows' conclusion. The reasoner chooses the remaining options with probabilities determined according to the atmosphere hypothesis. This hypothesis proposes that a conclusion should fit the premises' 'atmosphere', namely, the sentence types of the premises [1]. In particular, whenever at least one premise is negative, the most likely conclusion should be negative; whenever at least one premise contains 'some', the most likely conclusion should contain 'some' as well; otherwise the conclusion are likely to be affirmative and universal. Formally, the probability that the reasoner will switch to the guessing model is given by:

$$\frac{w_G}{w_G + \sum_{r' \in R} c_{r'} \cdot w_{r'}}$$

There are five possible outcomes of the guessing scenario: the subject could guess any conclusion, or could decide that nothing follows from the premises. The probability of nothing follows, given that the guessing scenario is chosen on the previous step, is computed as

$$\frac{v_{dl}}{3 \cdot v_{nd} + v_{as} + v_{dl}},$$

where $v_{dl} = \frac{1}{u_{t_1} + u_{t_2}}$ quantifies doubts of the reasoner that any valid conclusion can be derived from the premises. The quantity v_{dl} is computed relying on the amount of informativeness of both premise sentences (see [4]), the informativeness parameters u_t are estimated from the data and depend on the type t of a sentence (A, I, E or O). In the above expression, t_1 and t_2 refer to the types of sentences in the premises. The probability of guessing the conclusion predicted by the atmosphere hypothesis is:

$$\frac{v_{as}}{3 \cdot v_{nd} + v_{as} + v_{dl}},$$

where v_{as} is the weight assigned to the atmospheric hypothesis (also estimated from the data). Finally, the probability of guessing any of the remaining three options is

$$\frac{v_{nd}}{3 \cdot v_{nd} + v_{as} + v_{dl}},$$

where v_{nd} is a model parameter.[4]

---

[4]Without loss of generality, we set it to 1 as the model is over-parameterized.

Predictions \ Exp. Data	< 30%	≥ 30%
< 30%	Correct Rejection	Miss
≥ 30%	False Alarm	Hit

Table 2: Break-down of Predictions

The probability that a subject could arrive at a particular syllogistic conclusion is estimated from the tree by summing over all the leaf nodes containing the conclusion. Consequently, we can obtain posterior distribution of conclusions given the premises. These posterior distributions (for each premises) can be treated as model predictions, and we evaluate them (on unseen test set) against the distribution of human conclusions.

2.4 Estimation

We use the data from the meta-analysis by Chater and Oaksford [3], as is shown in Table 1. We denote the dataset as $\{X_i, y_i\}_{i \leq n}$, where X_i stands for the pair of premises and y_i stands for the conclusion. We randomly select 50% of the premises (i.e., half of the dataset) and use the corresponding examples as the training data. The rest of the data is used for evaluation. We use maximum likelihood estimation to obtain the parameter values. As our derivations are latent, there is no closed form solution for the optimization problem. Instead we use a variant of the Expectation Maximization (EM) algorithm which starts with a randomly initialized model and alternates between predicting derivations according to the current model (E-step) and updating model parameters based on these predictions (M-step, maximization of the expected likelihood). In our approach, the set of potentially applicable rules is determined by the reasoner state and, consequently, this set is not constant across the states (as discussed above, c_r was dependent on the state S). This implies that, unlike standard applications of EM, there is no closed-form solution for the M-step of the algorithm. Instead we use so-called generalized EM: instead of finding a maximum of the expected likelihood at M-step, we perform just one step of stochastic gradient ascent.

3 Results and Discussion

3.1 Evaluation

We use a mixed means of evaluation. We mainly use the evaluation method proposed in [11], which is based on the signal detection theory. The authors assume that the conclusions of the participants are noisy, that is unsystematic errors occur frequently. Hence, they classify the experimental data into two categories: those conclusions that appear reliably more often than chance level, which a theory of the syllogisms should predict to occur; and those that do not occur reliably more than chance level, which a theory should predict will not occur. In our context, there are five possible conclusions that can be drawn by subject. The chance level is thus 20%. In the following, we count a conclusion as reliable if it is drawn significantly often, i.e., in at least 30% of the trials.[5] As far as a theory predicts what will be concluded from each pair of premises, the method can be applied to evaluate the theory. According to the type of fitting, the predictions of a model are classified into four categories, see Table 2.

---

[5]This is slightly different from what used by [11] since they also included the non-scholastic order syllogisms, hence there are nine possible conclusions in their experiments, while we have five.

Data Set	Correct Prediction		Size	Mean Entropy	
	Count	Percentage		Predictions	Data
Test Set	153	95.6%	160	0.901	0.875
Training Set	151	94.4%	160	0.830	0.852
Complete Set	304	95.0%	320	0.870	0.864
NVC Premises*	212	94.2%	225	0.939	0.921
Valid Syl. Premises*	92	96.8%	95	0.706	0.729
Valid Syllogisms	23	95.8%	24	N/ A	N/ A

Table 3: Predictions evaluated according to the [11] method.
* The NVC premises are those from which no valid conclusion follows; the valid syl. premises are those from which at least one valid conclusion follows.

Sentence Types	A	I	E	O
Informativeness parameters	1.11	0.33	0.19	-0.78

Table 4: Values of the Informativeness Parameters.

3.2 Results

Table 3 shows the results. We see that the model is doing a good job, its proportion of correct predictions approximating a 95%.

3.3 Discussion

3.3.1 The Informativeness Parameters

The values of the informativeness parameters, as shown in Table 4, allow to make an interesting observation. Recall that we assumed that informativeness determines the confidence the reasoner has in the premises and, hence, the probability with which he concludes 'nothing follows'. We made no assumptions on 'which type of sentences' are more informative. The training results show that the amount of informativeness follow the order:

$$A(1.11) > E(0.33) > I(0.19) > O(-0.78),$$

which completely coincides with the proposal by Chater and Oaksford [3]. Besides, we see that sentence type 'O' is exceptionally uninformative, which also agrees with the authors' suggestion. The values of the informativenesses were learnt by the model. The result supports then the theory of Chater and Oaksford that the probabilistic validity plays an important role in human reasoning.

3.3.2 Parallel Comparison to Other Theories of the Syllogisms

We examined the predictions of a number of existing theories of the syllogistic reasoning. We were able to obtain the predictions of the PSYCOP model from Rips. The rest of the predictions were obtained from Table 7 in [11][6]. The results of the comparison are summarized is shown in Table 5. As far as we can see from the presented data our model outperforms other models.

---

[6]The table provided the predictions of the syllogistic theories on both the syllogisms that follow the scholastic order and the ones that do not. Our data are restricted to the scholastic order. The restriction has no influence

Theory	Hit	Miss	False Alarm	Correct Rejection	Correct Predictions
Atmosphere	44	41	20	215	259 /80.9%
Matching	41	44	55	180	221 /69.1%
Conversion	52	33	12	223	275 /85.9%
PHM*	40	45	63	172	212 /66.3%
PSYCOP	45	40	26	209	254 /79.4%
Verbal Models*	54	31	29	206	260 /81.2%
Mental Models*	85	0	55	180	265 /82.8%
Ver. 1 Test Data	26	15	12	107	133/83.1%
Ver. 2 Test Data	33	8	3	116	149/93.1%
Ver. 3 Complete Data**	70	14	5	231	301/94.1%
Ver. 3 Test Data	37	4	3	116	153/95.6%

Table 5: Predictions of the Theories of Syllogisms: A Summary.
*: Due to the limitations of the data we were able to obtain, the corresponding theory is likely to perform better than what is shown in the table.
**: The data in this line result from a cross-test: we take the predictions on the test data, then switched the test data and the training data and train the model again to get the predictions on the other half of the data.

4 Conclusion and Future Work

We have developed a preliminary framework of combining Natural Logic and data-driven inference weights and applied it to model syllogistic reasoning. The computational model learns from the experimental data, and as a result it may represent individual differences and explains subjects' systematic mistakes. This is achieved by assigning weights to all possible inference rules using machine-learning techniques and available data. The system is based on a Natural Logic proof system by Geurts [6], but it is less arbitrary, since it is empirically informed. In our approach we specify a tendency parameter for each inference rule. The agent begins with a pair of syllogistic premises and adopts each possible inference rule with a certain probability. As a result the longer the proof the less likely it is that an agent will find it. This simple setting solves the logical omniscience problem: not all derivations are available. Moreover, the approach takes into account various cognitive factors. For instance, the model enables the agents to adopt illicit conversions (e.g., yielding 'All A are B' from 'All B are A') in order to explain some systematic errors. Other version includes heuristic guesses based on two psychologically grounded principles. Firstly, the probability of drawing certain conclusions depends on the informativeness of the premises. Secondly, the model relies on the atmosphere hypothesis, e.g., when there is a negation in the premises, the agent is likely to draw a negative conclusion. We implemented and trained the models using the methodology outlined above and the empirical data from Chater and Oaksford [3]. We used a generalized EM algorithm to estimate the model and used it to compute the most probable syllogistic conclusions. The model was evaluated using the detection theory methods proposed in [11] to assess the performance of the theories of syllogistic reasoning. The complete version of the model makes 95% correct predictions, and therefore, outperforms all other known theories of syllogistic reasoning. In conclusion, the proposed combination of ideas gives rise to new, improved models of reasoning, where Natural Logic has replaced abstract rules, and the probabilistic parameters were derived from the data.

---

on the predictions of the atmosphere, matching, and conversion theories. However, for the PHM, the verbal model theory, and the mental model theory, we are unsure about the consequences.

The syllogistic fragment is an informative yet small arena for theories of reasoning. A natural next step would be to extend the model to cover a broader fragment of natural language by exploring existing Natural Logics [7] and designing new logics. We should then study formal (e.g., computational complexity) and psychological (e.g., cognitive resources) properties of the obtained models to draw new psychological conclusions and test the models against the data. The Natural Logics are usually computationally very cheap [12]. This guarantees that our models will easily scale-up to natural language reasoning. The computational complexity analysis will allow assessing the resources and strategies required to perform the reasoning tasks, cf. [14]. This in turn should open new ways of comparing our approach with other frameworks in psychology of reasoning.

References

[1] Ian Begg and Peter Denny. Empirical reconciliation of atmosphere and conversion interpretations of syllogistic reasoning errors. *Journal of Experimental Psychology*, 81(2):351, 1969.

[2] Johan van Benthem. *Language in Action: Categories, Lambdas and Dynamic Logic*. North-Holland, Amsterdam & MIT Press, Cambridge, 1991.

[3] Nick Chater and Mike Oaksford. The probability heuristics model of syllogistic reasoning. *Cognitive Psychology*, 38(2):191–258, 1999.

[4] Nick Chater and Mike Oaksford. *The Probabilistic Mind: Prospects for Bayesian Cognitive Science*. Oxford University Press, 2008.

[5] Jakub Dotlačil, Jakub Szymanik, and Marcin Zajenkowski. Probabilistic semantic automata in the verification of quantified statements. In *Proceedings of the 36th Annual Conference of the Cognitive Science Society*, pages 1778–1783, 2014.

[6] Bart Geurts. Reasoning with quantifiers. *Cognition*, 86(3):223–251, 2003.

[7] Thomas Icard III and Lawrence Moss. Recent progress in monotonicity. *Linguistic Issues in Language Technology*, 9, 2014.

[8] Alistair Isaac, Jakub Szymanik, and Rineke Verbrugge. Logic and complexity in cognitive science. In Alexandru Baltag and Sonja Smets, editors, *Johan van Benthem on Logic and Information Dynamics*, volume 5 of *Outstanding Contributions to Logic*, pages 787–824. Springer International Publishing, 2014.

[9] Philip Johnson-Laird. An end to the controversy? A reply to Rips. *Minds and Machines*, 7(3):425–432, 1997.

[10] Philip Johnson-Laird and Ruth Byrne. *Deduction*. Lawrence Erlbaum Associates, Inc, 1991.

[11] Sangeet Khemlani and Philip Johnson-Laird. Theories of the syllogism: A meta-analysis. *Psychological Bulletin*, 138(3):427, 2012.

[12] Ian Pratt-Hartmann. Fragments of language. *Journal of Logic, Language and Information*, 13(2):207–223, 2004.

[13] Lance Rips. *The psychology of proof: Deductive reasoning in human thinking*. MIT Press, 1994.

[14] Jakub Szymanik. *Quantifiers and Cognition. Logical and Computational Perspectives*. Studies in Linguistics and Philosophy. Springer, forthcoming, 2016.

[15] Víctor Manuel Sánchez Valencia. *Studies on Natural Logic and Categorial Grammar*. PhD thesis, University of Amsterdam, 1991.