Improving Cognitive Models for Syllogistic Reasoning

Jonas Bischofberger (JonasBischofberger@web.de)
Cognitive Computation Lab, University of Freiburg, Germany

Marco Ragni (ragni@cs.uni-freiburg.de)
Cognitive Computation Lab, University of Freiburg, Germany

Abstract

Multiple cognitive theories make conflicting explainations for human reasoning on syllogistic problems. The evaluation and comparison of these theories can be performed by conceiving them as predictive models. Model evaluation often employs static sets of predictions rather than full implementations of the theories. However, most theories predict different responses depending on the state of their internal parameters. Disregarding the theories' capabilities to adapt parameters to different reasoners leads to an incomplete picture of their predictive power. This article provides parameterized algorithmic formalizations and implementations of some syllogistic theories regarding the syllogistic single-response task. Evaluations reveal a substantial improvement for most cognitive theories being made adaptive over their original static predictions. The best performing implementations are PHM, mReasoner and Verbal Models, which almost reach the MFA benchmark. The results show that there exist heuristic and model-based theories which are able to capture a large portion of the patterns in syllogistic reasoning data.

Keywords: syllogistic reasoning; cognitive modeling; model evaluation

Introduction

Syllogisms are an extensively studied problem in human cognition (for a discussion see, e.g., Khemlani & Johnson-Laird, 2012). An example of a syllogism is:

No diver is a waiter.
Some waiters are artists.

Therefore: Some artists are not divers.

A classical syllogism consists of two premises: In the example above *'No diver is a waiter'* and *'Some waiters are divers'*. When asked to find logically valid conclusions to syllogistic premises, subjects often respond with erroneous conclusions (e.g., Begg & Denny, 1969). For example, a common but invalid response to the syllogism above would be *'No diver is an artist'*.

During the last century, many cognitive theories have been developed to explain human reasoning on syllogistic reasoning tasks. However, none of them is yet universally accepted (Khemlani & Johnson-Laird, 2012). One reason for this is that the predictions of these theories seem to fail to capture a large portion of experimental data (Riesterer, Brand, & Ragni, 2019). However, such evaluations usually rely on static sets of predictions without taking the adaptive capabilities of the theories into account. They provide only a lower bound of the performance while the full predictive capabilities of syllogistic theories remains uncertain.

An obstacle toward a more comprehensive evaluation is that most syllogistic theories do not readily constitute a full *cognitive prediction model*, i.e., a model that is able to generate a prediction for individual participants. For this article, seven prevailing syllogistic theories have been formalized and implemented in the programming language Python to generate a response for each syllogistic reasoning problem. This allows an evaluation that respects the models' ability to adapt to individual reasoners, drawing a clearer picture of their possible predictive performance.

Background

A syllogism is a kind of logical argument that contains two premises, each premise shares one term while the remaining two are distinct. The terms in the premises (the sets of entities) can be arranged in four different ways called figures (Johnson-Laird & Steedman, 1978). The two premises each have one of four quantifiers, called moods, in their syllogistic combination: All (abbreviated by A), Some (I), None (E), and Some ... not (O). There are a total of 64 kinds of syllogisms disregarding differences in content. When terms like *diver* or *waiters* are replaced by generic terms like *A*, *B* and *C*, we can rewrite a syllogism by

All A are B.
Some B are C.

A well-formed conclusion relates the two non-shared properties *A* and *C*, for example *Some A are C*. Four different kinds of syllogistic propositions and two ways to order the terms *A* and *C* lead to eight possible conclusions to any syllogism. Together with the response NVC ('No valid conclusion'), this makes nine possible responses that can be given by a reasoner to a syllogism. In this article, we model the task of choosing exactly one out of these nine responses.

There are at least twelve theories that attempt to explain syllogistic reasoning. They can roughly be categorized into three domains (Khemlani & Johnson-Laird, 2012): *Heuristic theories* propose that conclusions are drawn quickly and intuitively, using apparent often syntactic features of the syllogism, such as the quantifiers (see below for an example). *Logic-based theories* propose a deliberate reasoning mechanism using formal inference rules on mental representa-

tions of syllogistic propositions, similar to logical deduction. *Model-based theories* propose a mental representation that corresponds to models, sets or diagrams. Reasoning takes place in the form of encoding, manipulating and drawing conclusions from the respective model-based representations.

In this work, the same theories have been implemented which Khemlani and Johnson-Laird (2012) provide a fixed set of predicted responses for: Atmosphere (Woodworth & Sells, 1935; Revlis, 1975), Matching (Wetherick & Gilhooly, 1995), Illicit Conversion (Chapman & Chapman, 1959; Revlis, 1975) and PHM (Chater & Oaksford, 1999) as heuristic theories, PSYCOP (Rips, 1994) as logic-based theory and Verbal Models (Polk & Newell, 1995) and Mental Models, including a classic implementation (Bucciarelli & Johnson-Laird, 1999) and mReasoner (Khemlani & Johnson-Laird, 2013), as model-based theories. This allows an exhaustive comparison of carefully collected fixed predictions with our adaptable implementations.

Not considered are the following five theories: Verbal Substitutions (Ford, 1995), Euler circles (Erickson, 1974), Venn diagrams (Newell, 1980), Monotonicity theory (Geurts, 2003) and Source-founding theory (Stenning & Yule, 1997), as they are not cognitive theories or did not provide a prediction for all 64 syllogisms.

Model Implementations

Every syllogistic reasoning model has to define a prediction function that maps each syllogism onto a subset of the nine possible responses. This prediction function depends on parameters which can be fitted to actual item-response pairs. A final prediction is obtained by choosing uniformly among the produced responses. This makes theories which produce more than one response compatible with a single-response task without the need to make additional model-specific assumptions.

Our implementations[1] are intended to be as faithful as possible to the original formulation of the theories. However, some theories do not unambiguously lead to a cognitive model. Some of them come with a full prediction model that can be used as reference, others may, for instance, only define some basic operations and leave the control structure open. So there is an inevitable degree of subjectivity in the implementation of some theories. For those theories that provide reliable and comprehensible reference predictions, these predictions could be reproduced.

The following sections outline a selected number of our implementations and the theories they are based on. Each implementation is visualized using a diagram which shows classes of mental representations (nodes) and operations (edges) that constitute mappings between these classes.

The Matching Theory

Matching theory (Wetherick & Gilhooly, 1995) is a simple heuristic theory that proposes that the quantifier of the conclusion corresponds to the quantifier of the premise that makes an assertion about the fewest entities. In effect, the authors' predictions follow from choosing among the premises of a syllogism according to the preference relation on the quantifiers:

$$No = Some\ not = Some > All$$

Turning the matching theory into a prediction model is straightforward: Simply produce all conclusions with the one or two quantifiers that follow from the preference relation applied to the premises at hand. For example, from the syllogism *All A are B. Some B are C* would follow *Some A are C* and *Some C are A*. While *Some A are B. No C are B* would entail *Some A are C*, *Some C are A*, *No A are C* and *No C are A*.

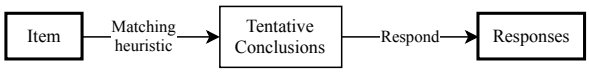

Figure 1: The Matching theory. Due to its single process of applying the matching heuristic no parameters can be used to adapt for differences between individual reasoners.

The capabilities of this simple one-operation model are limited: It can neither specify the term order of a conclusion, nor can it produce NVC, nor does it contain any adaptable parameters.

The Atmosphere theory (Woodworth & Sells, 1935) is similar to Matching in that it also consists of a single, non-adaptive heuristic applied to a syllogistic item. However, Atmosphere uses no preference relation over quantifiers but derives the quantifier of the conclusion from the combined atmosphere that the premise quantifiers are supposed to evoke.

The Probabilistic Heuristic Model (PHM)

A more complex and powerful heuristic theory is the Probability Heuristics Model by Chater and Oaksford (1999) which derives multiple heuristics from a probabilistic approach to reasoning. Centrally, they derive an order of informativeness I over the four proposition quantifiers:

$$I(All) > I(Some) > I(No) \gg I(Some\ not)$$

with $\gg$ much larger. The min-heuristic proposes that the quantifier of the most preferred conclusion corresponds to the quantifier of the least informative premise. The next most preferred conclusion quantifier follows from probabilistic entailment. For example *All A are C* probabilistically entails *Some A are C* because the conditional probability constraint $P(C \mid A) = 1$, corresponding to *All A are C*, entails that $P(C \mid A) > 0$ which corresponds to *Some A are C*.

While min-heuristic and p-entailment determine the quantifier of the conclusion, the attachment heuristic determines its term order. Attachment specifies that one of the two possible noun phrases (for example *Some A* or *Some C*) is chosen as subject noun phrase of the conclusion if it appears as subject noun phrase of one of the premises and the other one does not (Oaksford & Chater, 2001, p. 354). If both or none

[1] github.com/CognitiveComputationLab/cogmods/tree/master/syllogistic/2020_bischofsberger/

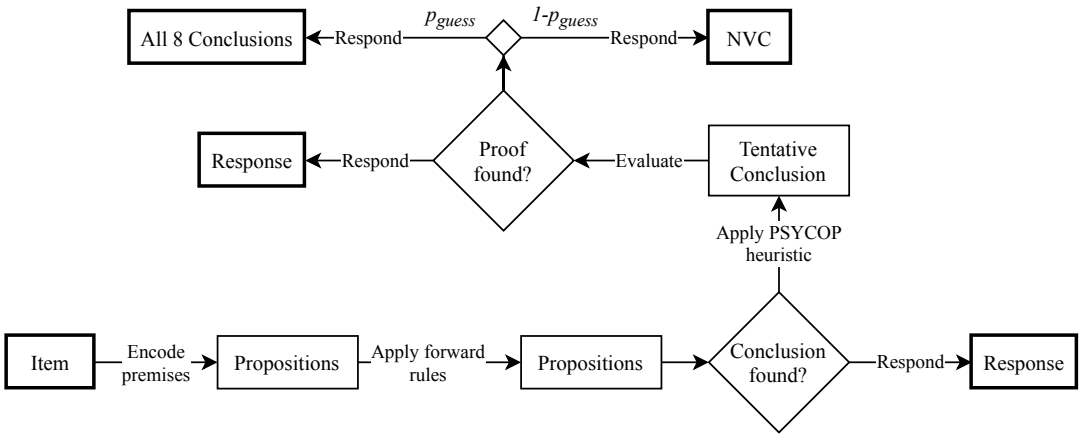

Figure 2: A conceptualization of the PSYCOP-inference process (based on the description by Rips, 1994, p. 244 ff).

of the candidate subject noun phrases appear in the premises, the end term order seems to be determined by using the end term of the most informative premise as subject. If the quantifiers of the premises are the same, this criterion fails and attachment yields no preferred term order.

Generated conclusions in PHM are evaluated using the max-heuristic, which proposes that the plausibility of a conclusion is proportional to the informativeness of the most informative premise.

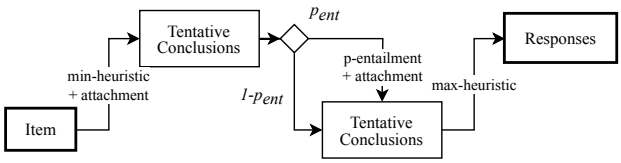

Figure 3: A conceptualization of the Probabilistic Heuristic Model (PHM) and possible adaptions for the individual reasoner.

The predictive model shown in Figure 3 includes parameters drawn from the original data description model (Chater & Oaksford, 1999, p. 212 f.). We augmented the original theory with five parameters. For example, the parameter p_{ent} that represents the amount of responses due to probabilistic entailments is converted to a parameter that corresponds to the probability that a probabilistic entailment is drawn after the min-heuristic has been applied. The max-heuristic is implemented using four parameters that determine the probability that a tentative conclusion is responded with, depending on the quantifier of the most informative premise. These confidence parameters are restricted to follow the order $p_{All} > p_{Some} > p_{No} > p_{Some\ not}$ according to the max-heuristic.

PSYCOP

PSYCOP (Rips, 1994) is a logic-based reasoning model which applies formal deduction rules to sets of encoded propositions in order to construct mental proofs. There are forward rules that start to operate on the premises and backward rules that start on a tentative conclusion. So the evaluation routine works from both directions. There are multiple parameters that control which deduction rules are available to the reasoner.

PSYCOP's deduction mechanism naturally applies to conclusion evaluation. To apply it to conclusion generation, the following routine has been proposed by Rips (1994, p. 244 ff.). First, apply all forward rules to the premises and if that yields a conclusion, respond with it. If no conclusion is found, a tentative conclusion is heuristically generated from the premises using the preference order

$$No > Some\ not > Some > All.$$

Then, the tentative conclusion is evaluated and responded with if a proof is found. If *no proof* is found, most participants are supposed to either guess or respond NVC. This structure has been implemented as shown in Figure 2. The implementation allows to control the different processes by twelve parameters (see the algorithm in the github repository).

A strong limitation of PSYCOP as a conclusion generation model is that its conclusion proposal heuristic only derives a quantifier but no term order from the premises. To avoid making additional assumptions, a term order is selected randomly in our implementation. Thus, PSYCOP's ability to capture patterns in the choice of term order seems highly limited.

Other than PSYCOP, the theory of Illicit Conversion (Chapman & Chapman, 1959) does not propose an explicit deduction process. Its main idea is that the mental representation of the premises is augmented with the converse of one or both premises. Our implementation of Illicit Conversion delivers such a possibly converted representation as input for a deduction process.

Mental Models

The theory of mental models (MMT for short) (Johnson-Laird & Steedman, 1978; Bucciarelli & Johnson-Laird, 1999) proposes that syllogisms are encoded as a mental representation of a situation in which the premises are true. For example, a mental model of the syllogism *Some A are B. Some B are not C.* might be:

```
a   b
a   b   -c
    b
```

This model contains three individuals. One with the properties A and B, one with the properties A, B and not-C and one with the property B. Conclusions can be either directly drawn from or evaluated based on a mental model. For example, the above model would entail the conclusion *Some A are not C*.

Central to MMT is the search for alternative models to refute tentative conclusions, i.e. counterexamples. While some reasoners may reply with their initial conclusion, others may try to refute it by imagining situations in which the premises are true but their tentative conclusion is false.

Two existing implementations of MMT have been reimplemented to compare them with the other models. The first one is a previous version (Bucciarelli & Johnson-Laird, 1999), with a LISP code which can be found online[2]. This implementation builds a mental model according to specific strategies from a syllogism and generates a conclusion per term order from this initial model. Afterwards, it may or may not generate counterexamples to its conclusions and update them by drawing new conclusions from the counterexample.

As shown in Figure 5, two parameters have been added to make this model fit a single-response task: p_{first} is the probability that an initial counterexample is searched for and $p_{further}$ holds the probability that an additional counterexample is searched for after at least one has already been found. Thus, the model can differentiate between reasoners who do not search for counterexamples at all, those who search for one counterexample and then stop and those who search for multiple counterexamples.

mReasoner

A more recent implementation of the MMT is mReasoner (Khemlani & Johnson-Laird, 2013). Its source code is also publicly available[3]. mReasoner uses a more powerful, stochastic, parameterized operation to build an initial mental model. It generates a set of initial conclusions from its mental model and with heuristics. According to this heuristic, the quantifier is determined by a preference order that is analogous to the order of informativeness in PHM. The end term order is determined by complex rules using both the quantifiers and the term order of the premises. mReasoner drops any proposed conclusion that does not hold in the initial mental model.

As the classical MMT, mReasoner either directly responds with its heuristic conclusions or tries to refute them, depending on a parameter p_λ. An additional parameter, p_ω, controls whether falsified conclusions are dropped entirely or weakened. In the latter case, mReasoner tries to refute the weakened conclusion again. The flow structure is shown in Figure 4.

mReasoner differs from the other implemented theories as it not only uses probabilities to string together different deterministic operations but its operations themselves may be stochastic. This is the case for its mental model encoding operation. Its high degree of stochasticity makes it especially suited to fit aggregated data.

While reasoners seem to use different strategies to find counterexamples (Bucciarelli & Johnson-Laird, 1999, p. 277), both implementations of the MMT do not contain parameters that encode a preference between different search strategies.

Another model-based theory, which resembles MMT, is the theory of Verbal Models (Polk & Newell, 1995). It is a highly adaptive theory that is focused on repeated encoding of the syllogism rather than the search for counterexamples.

The Evaluation of the Cognitive Models

All models are evaluated using the CCOBRA framework[4]. The repository provides two datasets, the `Veser2018` dataset (2058 items over 33 participants) is used for training and the `Ragni2016` dataset (8896 items over 139 participants) for evaluation.

The evaluation of an adaptive model works as follows: First, the model is pre-trained on the training set. For each parameter configuration of the model, an error is computed by comparing the predictions of the model using that parameter configuration with the actual responses in the training data. The parameter configuration with the lowest error is used as a starting point for the evaluation of each participant.

For the actual evaluation, CCOBRA traverses the test dataset participant by participant and then item by item. After the model has made a prediction for an item, its parameters are fitted to the actual response. After all items of a participant have been predicted, the model is reset to its state after pre-training and the evaluation process continues with the next individual in the test data. In the end, the overall predictive accuracy of the model corresponds to the fraction of correct predictions over the entire test dataset.

For comparison, the static reference predictions from Khemlani and Johnson-Laird (2012, Table 7) are turned into prediction models by employing the parametrizations (see above). Additionally, two benchmark models have been implemented: A Uniform Model that uniformly chooses one of the nine responses and the MFA model which

[2] `mentalmodels.princeton.edu/programs/Syllog-Public.lisp`

[3] `mentalmodels.princeton.edu/models/mreasoner/`

[4] `github.com/CognitiveComputationLab/CCOBRA`

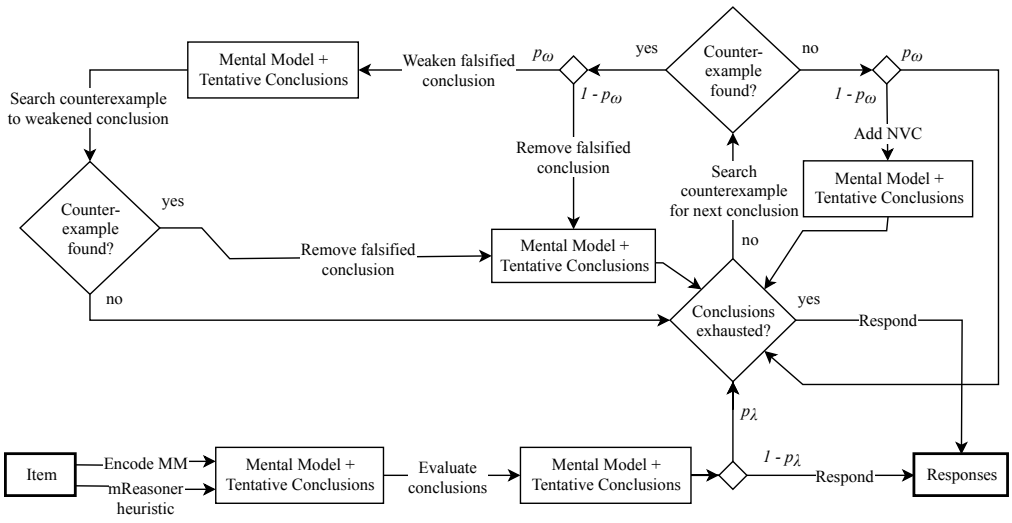

Figure 4: A conceptualization of the implementation of mReasoner (Khemlani & Johnson-Laird, 2013).

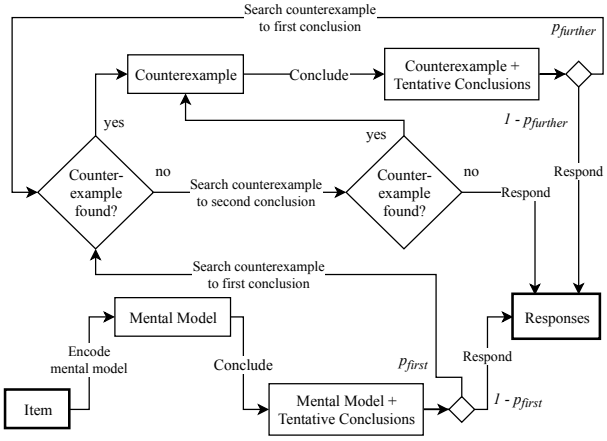

Figure 5: A conceptualization of the implementation of the classical MMT based on Bucciarelli and Johnson-Laird (1999) and possible parameter adaptations to predict the individual reasoner.

predicts the most frequent response given by participants to each syllogism in the training data available in CCOBRA.

Results and General Discussion

Figure 6 summarizes the predictive power of each implemented cognitive theory. Each model is more predictive than the random guessing model and thus is able to capture some amount of signal in the data. Also, the performance of every model except the Atmosphere theory, which does not use any parameters, could be substantially improved compared to the static predictions provided by Khemlani and Johnson-Laird (2012, Table 7). Some amount of this improvement can be explained without taking adaption into account. This is the case for the Matching theory, where the predictions from Khemlani and Johnson-Laird (2012, Table 7) differ from the original predictions (Wetherick & Gilhooly, 1995, Table 1) which were used for our implementation. But the bulk of the improvement likely stems from adaption.

The custom implementations can roughly be categorized into three levels of performance: The weakest ones are Atmosphere and Matching. They predict less than 25 percent of the responses correctly, which is likely due to their limiting simplicity and lack of adaptability. They generate their predictions according to simple mood-generating heuristics, which, in particular, cannot lead to the prediction of NVC. Consequently, they make wrong predictions for every item with a NVC response.

In the middle range we find PSYCOP, classic MMT and Illicit Conversion, indicating some possibilities for improving the predictive rate per participants. As mentioned, PSYCOP has been designed for conclusion evaluation rather than conclusion generation. It has a powerful conclusion evaluation routine but a weak, non-parameterized conclusion generation mechanism. With a more flexible way to propose conclusions for evaluation, it might be able to perform significantly better than the current version. The classical version of MMT has already been improved in the form of mReasoner outperforming its predecessor.

The best performing models are PHM, mReasoner and Verbal Models. Our Verbal Models implementation has the largest parameter space of all considered models, making it very adaptable. Also, it scores highest among the predictions from Khemlani and Johnson-Laird (2012, Table 7), indicating that it also provides a good baseline without considering individual adaption. However, its additional potential seems limited because of its large parameter space. It should also be noted that the verification of Verbal Models on the basis of the authors' original work proved to be difficult. Thus, a considerable degree of subjective assumptions had to be made for implementation. A different implementation may be able to

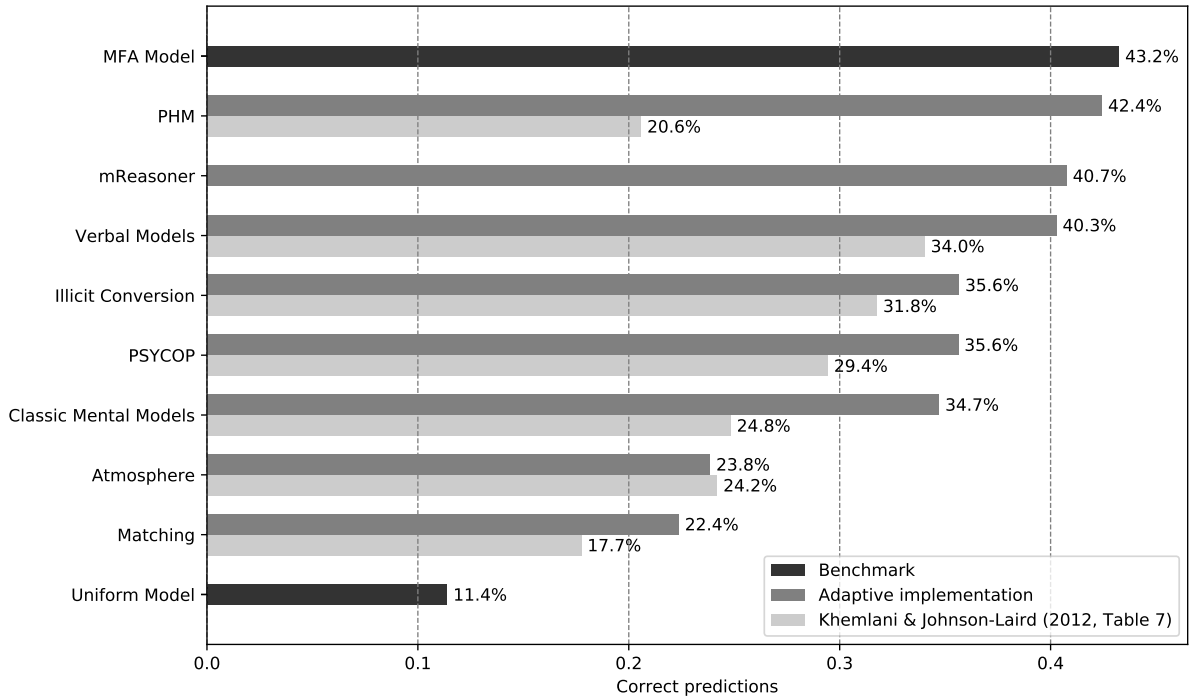

Figure 6: The predictive accuracies of the static predictions of the classical theories/models (as given in Khemlani & Johnson-Laird, 2012) and the improvements by our adaptive implementations.

lead to an improvement in predictive performance.

mReasoner has previously been shown to predict patterns in aggregated data using a stochastic prediction function (Khemlani & Johnson-Laird, 2016). Its success in predicting a single-response task demonstrates its general strength. Yet, there may remain some adaptive potential which could be exploited by controlling the choice of strategies for counterexample search via a parameter. Additionally, choosing between different plausible heuristics to generate conclusions may be beneficial for dual-process models like mReasoner.

PHM can be made a highly adaptive model by introducing parameters. It forms its prediction purely heuristically. It comprises a powerful set of heuristics to propose and evaluate syllogistic conclusions. Its high accuracy shows that a heuristic theory can reach the same level of performance as the best model-based theories on a syllogistic single-response task.

The evaluation shows that there are syllogistic models that are able to predict on average more than 40 percent of the conclusions drawn by an individual participant, almost reaching the level of the most frequent answer give (the MFA benchmark). Riesterer et al. (2019) identified an empirical upper bound for the performance of syllogistic prediction models by fitting various neural networks. Their results suggest an upper bound of roughly 50 percent, which is not far beyond the MFA benchmark. Thus, the best implemented models seem to be able to capture the majority of structure in the data.

Seven theories of syllogistic reasoning have been implemented as adaptive prediction models. Allowing the adaption to individual reasoners provides a significant gain in predictive performance for most models. While simple, non-adaptive heuristic models like Atmosphere and Matching seem not powerful enough to predict a large portion of responses correctly, there are models like PSYCOP and the classical MMT implementation which leave room for improvement or have already been improved. The potential for algorithmic improvements has been pointed out for some models. The best models PHM, mReasoner and Verbal Models perform close to the MFA benchmark and likely not too far below the theoretical upper bound. These results show that allowing individual adaption allows existing model-based and heuristic theories to explain the majority of responses given in syllogistic reasoning.

Acknowledgments

This paper has been partially supported by DFG grants RA1934/3-1, RA1934/4-1, and RA1934/9-1.

References

Begg, I., & Denny, J. P. (1969). Empirical reconciliation of atmosphere and conversion interpretations of syllogistic reasoning errors. *Journal of Experimental Psychology*, *81*(2), 351–354.

Bucciarelli, M., & Johnson-Laird, P. N. (1999). Strategies in syllogistic reasoning. *Cognitive Science*, *23*(3), 247–303.

Chapman, L. J., & Chapman, J. P. (1959). Atmosphere effect re-examined. *Journal of Experimental Psychology*, *58*(3), 220–226.

Chater, N., & Oaksford, M. (1999). The probability heuristics model of syllogistic reasoning. *Cognitive Psychology*, *38*(2), 191–258.

Erickson, J. R. (1974). A set analysis theory of behavior in formal syllogistic reasoning tasks.

Ford, M. (1995). Two modes of mental representation and problem solution in syllogistic reasoning. *Cognition*, *54*(1), 1–71.

Geurts, B. (2003). Reasoning with quantifiers. *Cognition*, *86*(3), 223–251.

Johnson-Laird, P. N., & Steedman, M. (1978). The psychology of syllogisms. *Cognitive Psychology*, *10*(1), 64–99.

Khemlani, S., & Johnson-Laird, P. N. (2012). Theories of the syllogism: A meta-analysis. *Psychological Bulletin*, *138*(3), 427–457.

Khemlani, S., & Johnson-Laird, P. N. (2013). The processes of inference. *Argument & Computation*, *4*(1), 4–20.

Khemlani, S., & Johnson-Laird, P. N. (2016). How people differ in syllogistic reasoning. In *Cogsci.*

Newell, A. (1980). Reasoning, problem solving and decision processes: The problem space as a fundamental category. In R. S. Nickerson (Ed.), *Attention and performance viii* (p. 693–718). Hillsdale, NJ: Lawrence Erlbaum Associates.

Oaksford, M., & Chater, N. (2001). The probabilistic approach to human reasoning. *Trends in Cognitive Sciences*, *5*(8), 349–357.

Polk, T. A., & Newell, A. (1995). Deduction as verbal reasoning. *Psychological Review*, *102*(3), 533–566.

Revlis, R. (1975). Two models of syllogistic reasoning: Feature selection and conversion. *Journal of Verbal Learning and Verbal Behavior*, *14*(2), 180–195.

Riesterer, N., Brand, D., & Ragni, M. (2019). Predictive modeling of individual human cognition: Upper bounds and a new perspective on performance..

Rips, L. J. (1994). *The psychology of proof.* MIT Press.

Stenning, K., & Yule, P. (1997). Image and language in human reasoning: A syllogistic illustration. *Cognitive Psychology*, *34*(2), 109–159.

Wetherick, N. E., & Gilhooly, K. J. (1995). 'atmosphere', matching, and logic in syllogistic reasoning. *Current Psychology*, *14*(3), 169–178.

Woodworth, R. S., & Sells, S. B. (1935). An atmosphere effect in formal syllogistic reasoning. *Journal of Experimental Psychology*, *18*(4), 451–460.