Necessity, Possibility and Likelihood in Syllogistic Reasoning

Daniel Brand (daniel.brand@metech.tu-chemnitz.de)
Predictive Analytics, Chemnitz University of Technology, Germany

Sara Todorovikj (sara.todorovikj@metech.tu-chemnitz.de)
Predictive Analytics, Chemnitz University of Technology, Germany

Marco Ragni (marco.ragni@hsw.tu-chemnitz.de)
Predictive Analytics, Chemnitz University of Technology, Germany

Abstract

In syllogistic reasoning research, humans are predominantly evaluated on their capabilities to judge whether a conclusion *necessarily* follows from a set of premises. To tackle this limitation, we build on work by Evans, Handley, Harper, and Johnson-Laird (1999), and present two studies where we asked participants for *possible* and *likely* conclusions. Combined with previous data (containing *necessary*), we present a comprehensive dataset with responses for all syllogisms, offering individual patterns for all three argument types -– a first of its kind. We discovered that *likely* serves as a middle ground between *possible* and *necessary*, paving the way to further investigate biases and preferences. Generally, individuals were able to handle the different notions, yet tended to interpret quantifiers in a pragmatic way, overlooking logical implicatures. Finally, we tested mReasoner, an implementation of the Mental Model Theory, and concluded that it was not able to capture the patterns observed in our data.

Keywords: Syllogistic Reasoning; Possibility; Mental Model Theory; Cognitive Modeling

Introduction

Despite being one of the oldest domains in human reasoning research (e.g., Störring, 1908), syllogistic reasoning is still far from being fully understood. Most commonly, syllogistic reasoning is investigated using traditional syllogisms that consist of two quantified statements (premises) with the first-order logic quantifiers *All*, *Some*, *No*, and *Some not*, which interrelate three terms. The task is usually to conclude what would necessarily follow from those premises, like in the following example:

All A are B.
Some B are C.

What, if anything, follows?

Since most research revolves around the traditional syllogisms with the task of finding necessary conclusions, a large variety of theories and models exist that aim for explaining and accounting for the observed behavior (for an overview, see Khemlani & Johnson-Laird, 2012). However, the focus of the strictly logic-based task to find necessary conclusions only covers a single aspect of the human reasoning capability. Attempts to go beyond the strict structure that is given by first-order logic, or to avoid some of the occurring problems with pragmatic interpretations, syllogistic quantifiers are often extended to generalized quantifiers (e.g., Brand, Mittenbühler, & Ragni, 2022; Tessler & Goodman, 2014) or varied in order to avoid misunderstandings due to pragmatic interpretations (e.g., Johnson-Laird & Byrne, 1989; Schmidt & Thompson, 2008). However, another restriction remains untouched by this, since participants are still given the task to determine the necessity of conclusions only, although the ability to decide which conclusion is *possible* is important in everyday life (Ragni & Johnson-Laird, 2020). One of the few investigations was done by Evans et al. (1999), where participants were asked to decide whether conclusions were possible as well as necessary. Thereby, they uncovered consistent fallacies, that they could account for by an implementation of the mental model theory (e.g., Johnson-Laird, 1983). Still, those extensions remained a niche in the domain, especially if it comes to cognitive modeling. Out of the twelve accounts for syllogistic reasoning presented by Khemlani and Johnson-Laird (2012), the only openly available model able to predict the behavior for deciding if conclusions are *possible* is mReasoner (Khemlani & Johnson-Laird, 2013), an implementation of the Mental Model Theory (MMT; Johnson-Laird, 1983).

For the sake of space, syllogisms in the remainder of this paper are abbreviated following the notation used by Khemlani and Johnson-Laird (2012). The quantifiers are thereby abbreviated by letters (*All*: A, *No*: E, *Some*: I and *Some not*: O), and the order of the terms in the premises (also called "figure") is denoted according to the table below:

figure 1	figure 2	figure 3	figure 4
A-B	B-A	A-B	B-A
B-C	C-B	C-B	B-C

The syllogism in the example before would therefore be abbreviated by AI1. Responses are treated similarly by combining the quantifier with the direction of the conclusion (i.e., *Iac* for "Some A are C"). In cases where no conclusion was selected, we will refer to that as *None*.

In this paper, we aim at going another step in the direction of investigating related tasks in the syllogistic domain. To this end, we conducted two experiments, in which participants solved all 64 traditional syllogisms. However, like in the experiments by Evans et al. (1999), the first experiment focuses on the question what conclusions are *possible*. In the second experiment, the scope is extended further and participants are instead asked to select all conclusions that they consider to be *likely* given the premises. In terms of its logical interpre-

In L. K. Samuelson, S. L. Frank, M. Toneva, A. Mackey, & E. Hazeltine (Eds.), *Proceedings of the 46th Annual Conference of the Cognitive Science Society.*

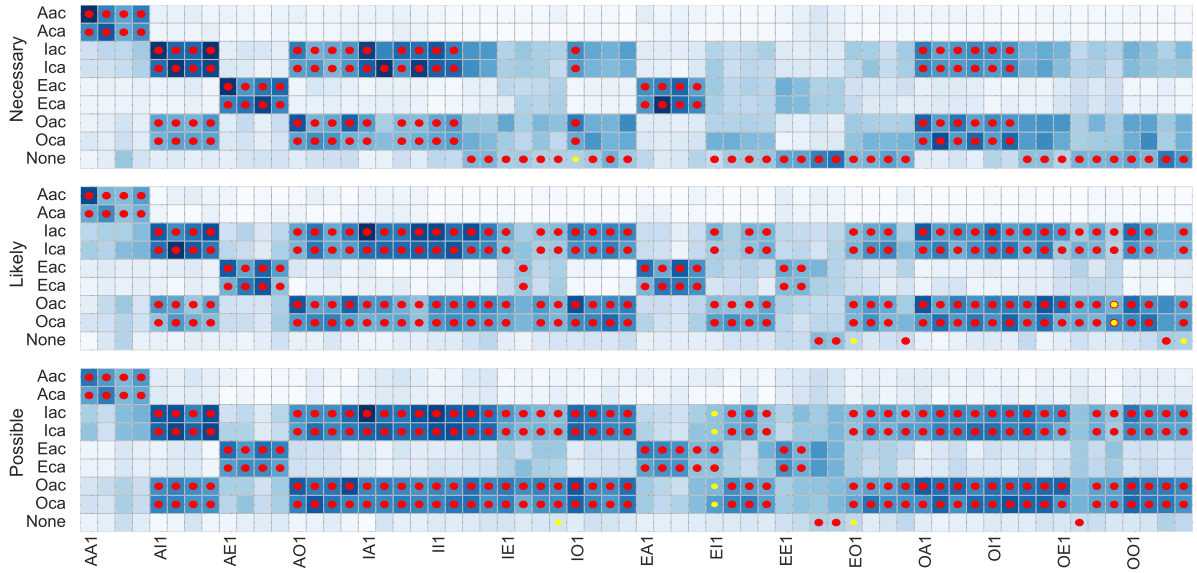

Figure 1: Response distributions for concluding *necessary*, *likely* and *possible* for all 64 syllogisms and 9 response options in our data. Darker shades of blue denote a higher proportion of the respective response option. Red circles denote the most frequently selected combination of response options (column-wise), yellow circles are used in case of a tie for the alternative.

tation, any subset of the *possible* conclusions that contains all *necessary* conclusions can qualify as *likely*, since no additional information is available. Therefore, asking for likely conclusions, in contrast to possible or necessary conclusions, can help to assess interpretations, preferences and biases, as it does not impose strict logical constraints on its own. Furthermore, we argue that the concepts of necessity and likelihood are, despite their different logical meaning, closely related in many instances of everyday reasoning. Since many situations do not require an exact assessment of conclusions, a reasonable estimate of necessity is sufficient. Therefore, we expect the behavior when asking for *likely* conclusions to be a "middle ground" between possibility and necessity.

The present article is structured as follows: First, we will present our studies and the datasets used. Second, we investigate our dataset thoroughly, with a focus on differences and similarities between the different task types and compare it with the dataset published by Evans et al. (1999). Third, we evaluate the capabilities of mReasoner to account for the observed behavior for both, necessary and possible. Finally, we discuss the results and the implications for syllogistic reasoning research as a whole.

Datasets

Experiment and Data Collection

For this work, we conducted two experiments on the online platform Prolific[1]. In both experiments, participants were asked to solve all 64 traditional syllogisms. The tasks had a multiple-choice design, where participants had to select all conclusions from a list of 9 options (8 combinations of quantifiers and direction, as well as the *None* option, to make it explicit that they do not want to select any option). Additionally, participants were asked if, according to their understanding, *Some A are B* also includes the possibility that *All A are B*. This was done since there are known problems with the interpretation of traditional quantifiers (e.g., Ceraso & Provitera, 1971), which lead to the suggestion to use alternative formulations (e.g., for *some*, as suggested by Schmidt and Thompson (2008) to avoid pragmatic responses) when investigating logical reasoning. However, in this work, we kept the traditional quantifiers to ensure comparability with existing datasets, in particular the dataset by Evans et al. (1999), as well as compatibility to models built based on traditional data. Additionally, it opens up the potential for investigations of pragmatic interpretations (e.g., based on the gricean maxim of quantity; Grice, 1975). Before the experiment ended, participants also had to solve the 7 question version of the Cognitive Reflection Task (Toplak, West, & Stanovich, 2014), which was found to be a predictor for syllogistic reasoning behavior (e.g., Brand, Riesterer, & Ragni, 2023).

The first experiment considered the task type *possible*, thus the participants were asked to select all conclusions that were possible given the premises. In total, the data of 50 participants was collected (19 female, 31 male). Only 13 (26%) of the participants stated that *Some* could also mean *All*. In the second experiment, the data of 49 participants (20 female, 29 male) was collected. This time, participants were asked to select all conclusions that they would consider to be likely given the premises. They were explicitly instructed to use their intuitive understanding for their responses. Out of the 49 participants, 9 (18.4%) participants stated that *Some* could also mean *All*.

[1] https://www.prolific.com/

Table 1: RMSE and MFA congruency between the different datasets. Since individual data is not available for the dataset by Evans et al. (1999), it is not possible to derive an MFA or how often participants rejected all conclusions (i.e., *None*). Therefore, *None* responses were ignored for the dataset and the MFA congruency was omitted.

Dataset 1	Dataset 2	RMSE	MFA
Necessary	Likely	.093	.78
Necessary	Possible	.154	.765
Likely	Possible	.101	.94
Necessary (Evans)	Possible (Evans)	.256	-
Necessary (Evans)	Necessary	*.205*	-
Possible (Evans)	Possible	*.278*	-

In previous work, a dataset containing the responses of 100 participants was collected in a similar experiment on Prolific, which were asked to select all conclusions that *necessarily* follow from the premises (Brand & Ragni, 2023). For this work, we combined the datasets to a comprehensive dataset that allows to investigate the effects of the task type. The dataset presented in this paper is – to our knowledge – the first dataset for syllogistic reasoning that covers *necessary*, *possible*, and *likely* for all 64 syllogisms with full information about the individual responses. Therefore, it can serve as a foundation for modeling endeavours (e.g., Tessler, Tenenbaum, & Goodman, 2022) or pattern analysis (e.g., Brand et al., 2023). The full dataset as well as the scripts used for the analyses are openly available on GitHub[2]. For simplicity, the three task types will be referred to as Necessary, Possible and Likely throughout the paper.

Evans et al. (1999) dataset

As a comparison, we also included the dataset published by Evans et al. in our analyses. The published data originates from the second of three experiments and contains the percentage endorsement of conclusions for *necessary* and *possible* for all 64 syllogisms and the 8 quantified conclusions. In total, the responses of 120 participants were collected, however, they were split into the task types (*necessary* and *possible*) and the conclusion direction (*A-C* and *C-A*), leading to 30 participants in each of the four groups.

Data Analysis

For easier comparability, Figure 1 shows the patterns for Necessary, Likely and Possible next to each other. Thereby, darker shades of blue denote a higher proportion of selections of the respective conclusion by the participants. Additionally, the most frequent answer combinations selected by the participants are highlighted with a red circle (in case of ties, a yellow circle denotes the other combination). For example, for the syllogism AA1, the most common combination of responses was Aac and Aca, despite Aac being selected more often in general.

As a first step, we aim at comparing the overall patterns in the datasets. Therefore, we calculate the root mean square error (RMSE) between the datasets to assess their similarity. Additionally, we also compare the patterns based on the most frequently given answer combinations (MFA) only by calculating the matching percentages. Table 1 shows the pairwise comparison between Necessary, Possible and Likely, as well as a comparison with the dataset by Evans et al. (1999). However, since no individual responses were provided in the dataset by Evans et al., it is not possible to derive the percentage of *None* responses, i.e., the number of cases where a participant rejected all conclusions. Therefore, we ignored *None* for the comparison with our datasets, which makes the comparison less expressive. When considering the RMSE, it becomes apparent that Likely seems to be between Necessary and Possible, while when considering the MFA, Likely is substantially closer to Possible ($MFA = .94$), whereas the MFA congruency compared to Necessary ($MFA = .78$) is only slightly higher than the MFA congruency between Possible and Necessary ($MFA = .765$). An interesting finding for Likely occurs when considering the results of the cognitive reflection task (CRT). The MFA congruency of only the subset of participants with a CRT score above the median is substantially closer to Necessary ($MFA = .832$) than participants scoring lower in the CRT ($MFA = .736$). For the dataset by Evans et al., the MFA congruency could not be calculated, since the individual combinations were not available.

The first thing that becomes apparent when looking at the patterns is the difference for *None*, which is prominent for Necessary, but not for the other two patterns. The difference was significant between Necessary and Possible (Necessary: 23.3% vs Possible: 12.6%; Mann-Whitney U test: $U = 3388.0$, $p = .0003$), indicating that participants seem to generally understand the logical difference between those, since *None* can only be the logically correct response when considering necessity (while, in contrast, most conclusions are always possible). This is in line with the findings of Evans et al. (1999), where participants also generally endorsed conclusions more when asked for the possibility than for necessity. For Likely, the percentage was comparable to Possible (12.0%), which, however, is hard to interpret due to its purely intuitive nature.

In order to get a more detailed understanding of the correctness of the responses, we calculated the correctness for each participant's selection. Since participants could select multiple conclusions, we relied on metrics considering the true/false positives and negatives which have shown to be beneficial to provide a more differentiated picture (e.g., Khemlani & Johnson-Laird, 2012). Table 2 shows the mean values for Accuracy, Precision, Recall and Specificity (for an introduction, see Fawcett, 2006) for Necessary and Possible.

For Necessary, the accuracy is substantially higher than for Possible. At first, this seems counter-intuitive, since identi-

[2] https://github.com/brand-d/cogsci-2024-likely

Table 2: Logical correctness for *necessary* and *possible* in terms of accuracy, precision, recall and specificity for the participants as well as for the predictions by mReasoner (mR).

Dataset	Accuracy	Precision	Recall	Specificity
Necessary	.767	.286	.486	.803
Possible	.416	.981	.363	.925
Nec. (mR)	.834	.499	.987	.817
Pos. (mR)	.661	1.0	.626	1.0

fying possible conclusions only requires finding a single example that is in line with the premises and should be the easier task. However, when considering the recall, it becomes apparent that participants generally select few conclusions (mean number of selected conclusions: Necessary: 2.21, Possible: 3.08, and Likely: 2.31). This – especially for Possible – deviates substantially from the average number of logically correct conclusions (1.328 for Necessary, 7.25 for Possible). Here, the CRT also has an effect on this number for Possible: Participants with an above-median CRT score select 3.501 conclusions on average, while it is only 2.678 for participants with a lower CRT score. Despite that, the overall selection for Possible seems to be most severely affected by interpretation issues. Since Necessary contains many syllogisms where only few conclusions are valid, the accuracy is also boosted by the true negatives. Interestingly, the precision is still low, indicating that the few selections still contain a significant number of false positives. Typically in syllogistic reasoning, this is connected to difficulties with invalid syllogisms: participants tend to not respond with *None* (or *No valid conclusion*) as often as necessary (e.g., Riesterer, Brand, Dames, & Ragni, 2020). Despite the low precision though, the specificity shows that, false positives are not the main cause of error. For Possible, the precision is very high for finding possible conclusions, which is due to the fact that for many syllogisms, every conclusion is possible. Therefore, participants have almost no chance to generate false positives, which is also corroborated by the high specificity. However, recall and accuracy show that they miss many conclusions overall. Put together, it shows that the participants generally miss conclusions rather than select too many. A possible explanation can be found in the fact that most participants used a pragmatic interpretation of the quantifier *I* (i.e., that "Some" does not include "All"): while, for example, selecting *Aac* would automatically imply *Iac* as well as *Ica* for both, Necessary and Possible, participants often did not.

To investigate these *implicatures*, we assessed the common occurrence of quantifiers, which is shown in Table 3. The values denote the proportion of cases an implication of the type $Q_1 \implies Q_2$ is fulfilled, i.e., the proportion of cases in which, if Q_1 is part of the selection, Q_2 is also selected. For Necessary and Likely, the proportions confirm the expectations by the pragmatic interpretation: The proportions of both, $A \implies I$ and $E \implies O$ are relatively low, while *I* and *O* are much more likely to imply each other. Interestingly, for Possible, this is not the case and all proportions are substantially higher. Despite not reaching the high proportions for *I* and *O*, $A \implies I$ holds true for more than half of the cases. However, the opposite direction, $I \implies A$ and $O \implies E$, are still low, although they would be logically warranted in the Possible task. An explanation could be that "All" and "Some" are interpreted as mutually exclusive. Since necessary implies that all selected conclusions have to be valid at the same time, they are not selected together, while possible allows that the conclusions can be considered independently. Another interesting point of the analysis is that Likely closely resembles the proportions of Necessary, which corroborates the assumption that typical preferences and patterns could also show for Likely (despite the lack of logical constraints). For Possible, the CRT score could serve as a predictor: The proportion for $A \implies I$ (.779) and $E \implies O$ (.662) of participants with an above-median score in the CRT is substantially higher than for participants with a lower CRT score (.42 and .153, respectively). The analysis also highlights the importance of multiple-choice designs in the domain, since the commonly used single-response designs, where participants provide a single conclusion for a syllogism, would hide the whole issue of quantifier interpretation: Since only a single conclusion could be given, the most preferential conclusion would be selected with no information about what other conclusions would also be potential candidates.

Table 3: Common occurrences of quantifiers for *Necessary*, *Likely* and *Possible* as well as for the predictions by mReasoner (mR). Values reflect the percentage of cases in which the second quantifier (Q_2) is selected when the first quantifier (Q_1) was selected as a response.

Q_1	Q_2	Nec.	Lik.	Pos.	Nec. (mR)	Pos. (mR)
A	I	.232	.277	.605	.915	.972
I	A	.047	.033	.133	.550	.801
E	O	.088	.141	.461	.953	.998
O	E	.05	.067	.231	.556	.805
I	O	.583	.609	.792	.871	.995
O	I	.617	.643	.755	.680	.843

Finally, we investigated the presence of the figural effect (e.g., Dickstein, 1978; Johnson-Laird & Bara, 1984), which predicts a bias depending on the figure of the syllogism: For syllogisms with figure 1 and figure 2, it predicts a bias towards *ac* conclusions and *ca* conclusions, respectively. When inspecting the patterns (see Figure 1), a tendency towards the figural effect becomes visible (especially for Likely and figural), although the most frequent answer combinations are bi-directional. To quantify the effect, we compare the number of responses in line with the figural effect (*Fig*) with the number of responses contradicting the figural effect ($\neg Fig$) for all syllogisms with figure 1 or figure 2 (*None* is ignored). The results for all datasets can be seen in Table 4.

Table 4: Figural effect in the *necessary*, *likely* and *possible* dataset as well as in the dataset by Evans et al. (1999). *Fig* and $\neg Fig$ denote the mean number of responses in line/not in line with the figural effect (*ac* for figure 1 and *ca* for figure 2), respectively. Additionally, the difference and the results of a Mann-Whitney U test is shown.

Dataset	*Fig*	$\neg Fig$	Diff	U	p
Necessary	0.932	0.813	0.118	681.5	**.023**
Likely	1.118	0.968	0.15	784.5	**<.001**
Possible	1.386	1.316	0.07	630.5	.115
Nec. (Evans)	1.521	1.542	-0.021	462.5	.517
Pos. (Evans)	2.297	2.32	-0.023	490.5	.785

The figural effect is significant for both, Necessary and Likely, while significance was not reached in the Possible dataset. Thereby, the effect was stronger for Likely ($Fig - \neg Fig = .15$) than for Necessary ($Fig - \neg Fig = .118$), indicating that the figural effect is not only a bias occurring when reasoning logically, but also a preference effect. Since no logical constraint affected the direction of the conclusions for Likely, the presence of the effect also reflects a preferred understanding of the premises. In the dataset by Evans et al. (1999), however, the figural effect is not present. Here, the experimental design is likely the cause: For both task types, possible and necessary, participants were divided into groups, where one group was only given *ac* conclusions, while the other group decided on the *ca* conclusions. Therefore, there is no information about the figural effect *within person* in the dataset, and the inter-individual differences between the groups likely overshadow the effect. This highlights the importance of full/complete datasets containing individual information: Analyses and modeling endeavours going beyond the originally intended scope can benefit substantially from the availability of the additional information, while artefacts due to missing data are prevented.

Model Analysis

Evans et al. (1999) showed that the Mental Model Theory was able to account for several observed effects and provides an explanation for the differences in reasoning behavior observable when participants solved syllogistic tasks for possible conclusions compared to the traditional task of finding necessary conclusions. Determining whether a conclusion is possible requires the reasoners to verify that it is consistent with the given syllogistic premises. Necessity, on the other hand, requires that the conclusion is validated by confirming that it is consistent not only with the premises, but also with all other potential conclusions. This is reflected in the three deduction stages of the Mental Model Theory (Evans et al., 1999). First, reasoners construct a model from the syllogistic premises, then derive an initial conclusion (possible). Finally, they engage in a search for counterexamples that might invalidate the conclusion, and if none are found – they accept it (necessary). These results, however, remained on the level of explaining effects, and did not use the Mental Model Theory to account for the actual response patterns, especially when also considering individual reasoners. As a well-established implementation of the Mental Model Theory that was used to account for the complete response behavior of individuals (Riesterer, Brand, & Ragni, 2020), mReasoner (Khemlani & Johnson-Laird, 2013) seems to be well-suited for the task, since it supports querying for *necessary* and *possible*. In the following analysis, we use mReasoner to generate a dataset that is as close as possible to the patterns for Possible and Necessary in our dataset, by approximating each participant individually.

Following Brand, Riesterer, and Ragni (2021), we fitted mReasoner to each individual participant in the respective datasets. Given the participants' responses, we search for a model configuration that minimizes the prediction error (RMSE) when queried on *Is it necessary?* and *Is it possible?*, essentially reconstructing the original datasets. Given the stochastic nature of the model, we employed a repeated sampling, querying for four predictions (samples) that we used to approximate the expected result. Since querying the model leads to acceptance or rejection of a given quantified conclusion, that means that no *None* responses were given. Therefore, we interpret the other responses as probabilities and estimate *None* by assigning it the probability that none of the other options were selected. This was essential for Necessary, in order to account for *No valid conclusion* responses.

Analogous to our analysis above, we obtained the response patterns derived by mReasoner by aggregating the individual predictions for each participant in our dataset. Observing the distribution shown in Figure 2, a discrepancy can be immediately noticed between the responses obtained from mReasoner and the true patterns (Figure 1). The inaccuracies of the patterns is corroborated by the errors (Necessary RMSE = .212; Possible RMSE = .253).

When looking into the implicatures (see Table 3), one cause of the errors becomes apparent. For $A \implies I$, mReasoner predicted a proportion of .915 and .972, whereas the results for $E \implies O$ were even higher with .953 and .998 for Necessary and Possible, respectively. This indicates that mReasoner fails to capture the lack of logical behavior observed in the human responses. It shows a much closer connection to logical implicatures than to the pragmatic interpretation of *Some*.

This is also reflected in the measured logical correctness in Table 2. All of the values exceed the measured correctness of the participants, especially in the case of Necessary, where the recall value difference is particularly prominent (mReasoner: .987, original: .486). Regarding Possible, a closer resemblance may be observed, however mReasoner still overestimated the amount of logically correct responses. Overall, mReasoner did not manage to appropriately capture the participants' behavior in our dataset. The main source of error seems to originate in the assumption that participants are re-

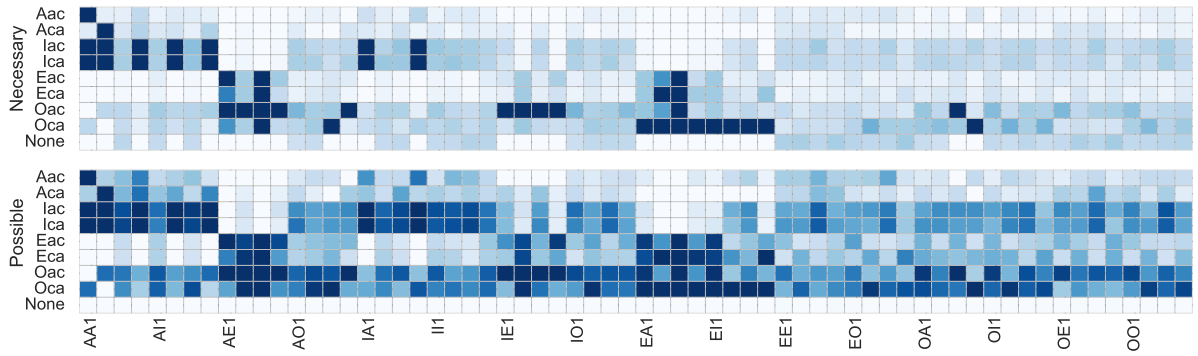

Figure 2: Response distributions for concluding *necessary* and *possible* obtained from mReasoner for all 64 syllogisms and 9 response options. Patterns were obtained by fitting mReasoner to each individual participant in the respective dataset using multiple samples for each conclusion. Darker shades of blue denote a higher proportion of the respective response option.

lying on the logical relations between quantifiers, while pragmatic interpretations were dominant in our dataset.

Finally, we tested mReasoner on the dataset by Evans et al. (1999), which we expected to improve the performance, since pragmatic interpretations seemed to be less prominent in the dataset. Since the dataset did not include individual responses, we were not able to apply the same fitting process. Instead, we fitted mReasoner to the dataset as a whole, searching for the best combination of 30 patterns based on the RMSE (since each group had 30 participants in the dataset). While the resulting RMSE was better (Necessary RMSE = .23; Possible RMSE = .196), it still did not reflect the response behavior well considering that the RMSE between possible and necessary in the dataset was only .256.

Discussion

We conducted two experiments, obtaining syllogistic reasoning data that altered the commonly used task to find necessary conclusions. Thereby, we extend the work by Evans et al. (1999), by not only including the question for possible, but also for likely conclusions. Combined with data from previous work containing Necessary (Brand & Ragni, 2023), we present a comprehensive dataset that contains multiple-choice responses for all 64 syllogisms and 9 response options from each participant, covering Necessary, Possible and Likely as a task type. The presented dataset is, to our knowledge, the first dataset that offers complete individual patterns for the three task types. Our analyses showed that participants generally seem to be able to grasp the differences between Possible and Necessary, corroborating the findings by Evans et al. (1999). This showed in distinct patterns for both task types and a significantly lower rate of responses that rejected all conclusions when asked for possibility.

Considering Likely, it became apparent that it serves as a middle ground between Possible and Necessary. While possibility still is the precondition for likelihood, there are no further logical restrictions implied with the task type. Therefore, it is well-suited to investigate biases and preferences, allowing to disentangle effects in human logical deduction from preferences in everyday reasoning. An example for this was the figural effect, which was not only present for Necessary, but instead was even stronger for Likely while it was not significant for Possible. This indicates that the figures of the premises are interpreted as implicitly hinting at some conclusions, making them appear more likely (even without any logical necessity). Our dataset also showed that most participants interpreted the quantifiers in a pragmatic way, instead of using first-order logic. When explicitly asked, most participants rejected the statement that *Some* would also include the possibility for *All*. In the resulting reasoning patterns, participants were in line with their stated interpretation, since *All* often did not imply *Some*, although it was logically warranted. Overall, the logical correctness seemed to be mostly influenced by these effects, highlighting the importance of investigating interpretations (e.g., Roberts, Newstead, & Griggs, 2001). When evaluating the capabilities of mReasoner, it became apparent that the logical correctness assumed by the model was too high. One reason was likely the aforementioned difference in interpretation, that was not accounted for by the model. However, even when compared to the dataset by Evans et al. (1999), which does not seem to be influenced as heavily by that, mReasoner could not capture the whole pattern sufficiently.

Finally, the issue with different interpretations of quantifiers poses an interesting question: When aiming at investigating the actual logical deduction in human reasoning, it might prove to be beneficial to replace, as suggested by Schmidt and Thompson (2008), the traditional quantifiers by less ambiguous ones. This, however, hinders the possibility to investigate and model effects originating from pragmatic reasoning and natural language (e.g., Tessler & Goodman, 2014; Tessler et al., 2022) and limit the comparability to the extensive data stock acquired over the long history of syllogistic reasoning research and the large variety of models built upon traditional syllogisms. To this end, we decided to stick with the traditional syllogistic structure, while aiming to incrementally widen its scope to investigate new facets in human syllogistic reasoning.

Acknowledgements

This project has been partially funded by a grant to MR in the DFG-projects 529624975 and 283135041.

References

Brand, D., Mittenbühler, M., & Ragni, M. (2022). Generalizing syllogistic reasoning: Extending syllogisms to general quantifiers. In J. Culbertson, A. Perfors, H. Rabagliati, & V. Ramenzoni (Eds.), *Proceedings of the 44th Annual Conference of the Cognitive Science Society* (pp. 722–728).

Brand, D., & Ragni, M. (2023). Effect of response format on syllogistic reasoning. In M. Goldwater, F. K. Anggoro, B. K. Hayes, & D. C. Ong (Eds.), *Proceedings of the 45th Annual Conference of the Cognitive Science Society* (pp. 2408–2414).

Brand, D., Riesterer, N., & Ragni, M. (2021). Unifying models for belief and syllogistic reasoning. In T. Fitch, H. Lamm, H. Leder, & K. Teßmar-Raible (Eds.), *Proceedings of the 43th Annual Meeting of the Cognitive Science Society* (pp. 2801–2807).

Brand, D., Riesterer, N., & Ragni, M. (2023). Uncovering iconic patterns of syllogistic reasoning: A clustering analysis. In C. Sibert (Ed.), *Proceedings of the 21th International Conference on Cognitive Modeling* (pp. 57–63). University Park, PA: Applied Cognitive Science Lab, Penn State.

Ceraso, J., & Provitera, A. (1971). Sources of error in syllogistic reasoning. *Cognitive Psychology*, *2*(4), 400-410. doi: 10.1016/0010-0285(71)90023-5

Dickstein, L. S. (1978). The effect of figure on syllogistic reasoning. *Memory & Cognition*, *6*, 76–83.

Evans, J., Handley, S., Harper, C., & Johnson-Laird, P. (1999). Reasoning about necessity and possibility: A test of the mental model theory of deduction. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *25*, 1495-1513. doi: 10.1037/0278-7393.25.6.1495

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters*, *27*(8), 861-874. doi: 10.1016/j.patrec.2005.10.010

Grice, H. P. (1975). Logic and conversation. In P. Cole & J. L. Morgan (Eds.), *Syntax and Semantics. Vol. 3 : Speech Acts* (pp. 41–58). New York: Academic Press.

Johnson-Laird, P. N. (1983). *Mental models: Towards a cognitive science of language, inference, and consciousness*. Cambridge, MA, USA: Harvard University Press.

Johnson-Laird, P. N., & Bara, B. G. (1984). Syllogistic inference. *Cognition*, *16*(1), 1–61.

Johnson-Laird, P. N., & Byrne, R. M. (1989). Only reasoning. *Journal of Memory and Language*, *28*(3), 313–330.

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

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

Ragni, M., & Johnson-Laird, P. N. (2020). Reasoning about epistemic possibilities. *Acta Psychologica*, *208*, 103081. doi: https://doi.org/10.1016/j.actpsy.2020.103081

Riesterer, N., Brand, D., Dames, H., & Ragni, M. (2020). Modeling human syllogistic reasoning: The role of "No Valid Conclusion". *Topics in Cognitive Science*, *12*(1), 446–459.

Riesterer, N., Brand, D., & Ragni, M. (2020). Do models capture individuals? Evaluating parameterized models for syllogistic reasoning. In S. Denison, M. Mack, Y. Xu, & B. C. Armstrong (Eds.), *Proceedings of the 42nd Annual Conference of the Cognitive Science Society* (pp. 3377–3383). Toronto, ON: Cognitive Science Society.

Roberts, M. J., Newstead, S. E., & Griggs, R. A. (2001). Quantifier interpretation and syllogistic reasoning. *Thinking & Reasoning*, *7*(2), 173-204. doi: 10.1080/13546780143000008

Schmidt, J., & Thompson, V. (2008). "at least one" problem with "some" formal reasoning paradigms. *Memory & cognition*, *36*, 217-29. doi: 10.3758/MC.36.1.217

Störring, G. (1908). *Experimentelle Untersuchungen über einfache Schlussprozesse*. W. Engelmann.

Tessler, M. H., & Goodman, N. D. (2014). Some arguments are probably valid: Syllogistic reasoning as communication. *Proceedings of the 36th Annual Conference of the Cognitive Science Society*, 1574-1579.

Tessler, M. H., Tenenbaum, J. B., & Goodman, N. D. (2022). Logic, probability, and pragmatics in syllogistic reasoning. *Topics in Cognitive Science*, *14*(3), 574-601. doi: 10.1111/tops.12593

Toplak, M. E., West, R. F., & Stanovich, K. E. (2014). Assessing miserly information processing: An expansion of the cognitive reflection test. *Thinking & Reasoning*, *20*(2), 147-168. doi: 10.1080/13546783.2013.844729