Model Verification and Preferred Mental Models in Syllogistic Reasoning

Sara Todorovikj (sara.todorovikj@hsw.tu-chemnitz.de)
Daniel Brand (daniel.brand@hsw.tu-chemnitz.de)
Marco Ragni (marco.ragni@hsw.tu-chemnitz.de)
Predictive Analytics, Chemnitz University of Technology
Straße der Nationen 62, 09111 Chemnitz, Germany

Abstract

A core cognitive ability of humans is the creation of and reasoning with mental models based on given information. When confronted with indeterminate information, allowing for the existence of multiple mental models, humans seem to recurrently report specific models - so-called preferred mental models. In this paper, we revisit this within the context of syllogistic reasoning, which involves statements about quantified assertions. We present an experiment designed to investigate the verification process of preferred mental models. Our analysis centers on two primary research questions: Is model verification generally straightforward for reasoners? And does a preference effect for specific models exist in syllogistic reasoning? Furthermore, employing modeling techniques, we analyze the structural complexity of mental models, based on the types of instances they consist of. We discuss our findings and their implications on the differences between reasoning with syllogisms and spatial statements.

Keywords: Mental Model Theory; Preferred Mental Models; Syllogistic Reasoning; Individual Differences.

Introduction

Consider the following reasoning example:

All blue shapes are circles.
All blue shapes have a diamond mark.

What, if anything, follows?

This problem is a so-called syllogism. The task at hand is to determine what kind of relation, if any, exists between the two end-terms, *circles* and *diamond (mark)*, also called subject and predicate, respectively. In general, a syllogism is defined by its quantifiers (*mood*) and term order (*figure*). We take into consideration these four first-order logic quantifiers: *All* (A), *Some* (I), *Some...not* (O) and *None* (E). The figure is determined by the order of the subject, middle term and predicate of the syllogism, represented by A, B and C, respectively, in the following notation (adopted from Khemlani & Johnson-Laird, 2012):

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

A syllogism can be denoted using the given mood abbreviations and figure, for example the syllogism above is AA4 Conclusions are denoted in a similar fashion using the quantifier's abbreviation and the order of the end-terms (*ac* or *ca*), e.g. *Eca* denotes 'No C are A'. Finally, 'No valid conclusion' is abbreviated by NVC. There exist at least twelve theories that aim to explain and model the processes behind human syllogistic reasoning (for an overview, see Khemlani & Johnson-Laird, 2012). One of the most prominent theories among them is the Mental Model Theory (MMT; e.g., Johnson-Laird, 1975, 2010). MMT postulates that given some observations, individuals create iconic representations – *mental models* – of possibilities. They create their own subjective mental representation of the information presented in a reasoning task. Considering the example above, one possible representation would be:

circles	[blue]	[diamond]
circles		

The square brackets around an instance denote that the set of entities described by it is exhaustively represented. Another possible mental model representation is:

circles		
circles	[blue]	diamond
¬circles	¬blue	diamond

where ¬ denotes negation. Both mental representations support the conclusion "Some circles have a diamond mark" - the logically valid conclusion to this syllogism. However, in order to confirm the validity, an individual should think of all possible premise interpretations and check if they hold. The expansion of the interpretation search space can make solving such problems difficult for humans (Johnson-Laird, 2008).

Preferred Mental Models

An empirical phenomenon has been reported in the literature concerning problem descriptions allowing for multiple possible models. Specifically, some models are preferred over others – such models are called preferred mental models (PMM).

Spatial Reasoning Spatial relational reasoning problems which can evoke multiple mental models, are not all created equally (Knauff, Rauh, & Schlieder, 1995; Ragni & Knauff, 2013). This has been demonstrated through model acceptance tasks, where participants were asked to decide whether a presented spatial arrangement matches a given set of indeterminate premises. Both the patterns of acceptance responses and the reaction times clearly show that some models are preferred over others and these models adhere to some simple construction principles.

Table 1: Canonical and non-canonical instances for a syllogistic premise with terms X and Y according to mReasoner (Khemlani et al., 2015), presented in Todorovikj et al. (2023)

Quantifier	Canonical	Non-canonical
All	X Y	¬X Y ¬X ¬Y
Some	X Y X ¬Y	¬X Y ¬X ¬Y
No	¬X Y X ¬Y	¬X ¬Y
Some not	X Y X ¬Y ¬X Y	¬X ¬Y

Syllogistic reasoning Todorovikj et al. (2023) investigated the model building process in syllogistic reasoning by empirically testing what kind of models individuals create when presented with syllogistic premises. They designed an experimental domain of objects described by their shape, colour and mark. The experiment they conducted presented participants with a syllogism describing such objects and prompted them to provide a visual representation of the premises by selecting and creating objects with their desired attributes. They found that 82% of the models were correct representations of the syllogism. After analyzing the response patterns and identifying the most frequent ones, the authors reported finding preferred mental models for 46 out of 64 syllogisms. Additionally, they examined whether the observed model building behavior is in line with the model building processes of *mReasoner*[1], a LISP-based implementation of the MMT (Khemlani & Johnson-Laird, 2013). During that analysis they did not find significant results that would confirm the relevance of the initially constructed models for the final conclusion, allowing for the possibility that the built models are not necessarily the ones used when reasoning.

Canonicality of Mental Models

In mathematical and computer sciences, canonicality refers to minimal representations that avoid redundancy and ambiguity while capturing the essential properties of an expression. Within the domain of mental models in syllogistic reasoning canonicality describes the necessity of possible instances (Khemlani et al., 2015). Specifically, which entities are absolutely necessary to represent a syllogism correctly (*canonical* set of instances), and which ones do not have to be present, but do not falsify the premises and therefore could possibly be included in a model (*non-canonical* set of instances). The canonical and non-canonical instances that can be used for building a model based on the LISP implementation of mReasoner are displayed in Table 1. When building a model in mReasoner, the ε parameter is used to describe the likelihood that an instance is drawn from the full set of possible instances in contrast to only the canonical one (Khemlani et al., 2015). When fitting the model to their data, Todorovikj et al. (2023) used the proportions of non-canonical instances in the model to approximate the respective ε value.

In this article, we reinforce the first definition of canonicality when we describe syllogistic models. We define a *canonical model* as the minimal representation of a syllogism and a *non-canonical model* as the opposite extreme, i.e., a maximal representation. For example, consider the syllogism AA1:

All squares are blue.
All blue shapes have a star mark.

Its canonical model would only consist of entities of the following instance:

[square] [blue] [star]

The non-canonical model on the other hand, would consist of all these instances (examples of negations in red):

[square]	blue	star
triangle	blue	star
triangle	red	star
triangle	red	cross

Analogously, we define an *incorrect canonical model* as the minimal incorrect representation and an *incorrect non-canonical model* as the maximal one. In the following experiment and analysis we will use these definitions of canonical and non-canonical models as lower and upper bounds of a model's complexity and heterogeneity.

Ultimately, we pose the following two research questions that we aim to answer in this paper:

[RQ1] Is the verification of models generally easy for reasoners? How fast and accurate is that process?

[RQ2] Do preference effects for accepting models in syllogistic reasoning exist? Are certain models more likely to be accepted or rejected correctly and faster than others?

The remainder of the paper is structured as follows: We first describe our experimental design, followed by an analysis of the participants' data. Afterwards, we go in-depth with respect to the structural properties of the models and outline a regression model based on them. We conclude with a discussion of our results.

Experiment

In the experiment we conducted, participants were shown a set of syllogistic premises, followed by a visual description of a model corresponding to the syllogism, which they were asked to accept or reject. Following Todorovikj et al. (2023), the syllogistic contents were object descriptions in terms of their *shape* (circle, triangle, square), *color* (red, yellow, blue) and *mark* (plus, star, diamond). We take into consideration only the 46 syllogisms for which a preferred mental model was found. For each one of them we created six tasks by deriving the preferred mental model (PMM), the canonical

[1] https://github.com/skhemlani/mReasoner

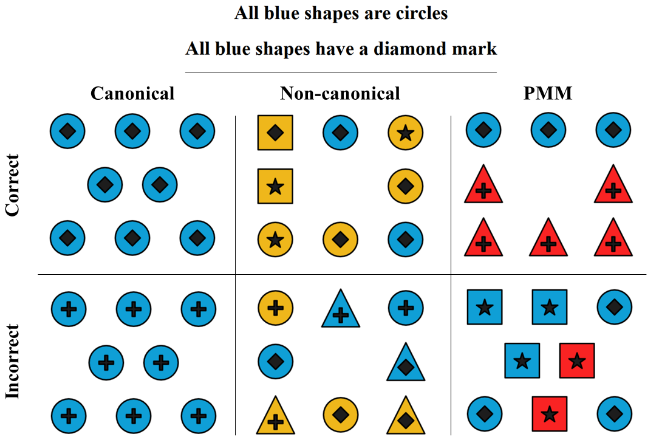
All blue shapes are circles
All blue shapes have a diamond mark
Canonical
Non-canonical
PMM
Correct
Incorrect

Figure 1: Illustrative example of the six models for the syllogism AA4 with contents *circle*, *blue* and *diamond*. A task in the experiment consists of two syllogistic premises describing properties of shapes and a visual representation of a specific model, as depicted here. Participants are asked to decide whether the model contradicts the premises.

model, the non-canonical model and an incorrect counterpart for each one of them, as control.

The PMMs were directly obtained from the experimental data of Todorovikj et al. (2023). For the other two models we obtained the (non-)canonical instances for both quantifiers, following Table 1, merged them based on the middle term when possible, while ensuring that they do not falsify the premises. If a merge is not possible when deriving the canonical model, then a non-canonical instance is introduced. In the case of multiple potential merges, one that minimizes negativity is chosen. That is motivated by the *principle of truth* (Johnson-Laird, 1983, 2008) which states that mental models are constructed to represent what is true according to the reasoner. We illustrate this process taking the syllogism AI2 and its canonical model as an example:

All B are A.   Some C are B.

The canonical instances for each premise are:

$$\text{A B} \qquad \begin{array}{c} \text{B C} \\ \neg\text{B C} \end{array}$$

The instance AB can be immediately merged with BC into ABC, which does not falsify the premises. For ¬BC, we look into the non-canonical set of the first premise, which contains A¬B and ¬A¬B, both eligible to merge with ¬BC, making A¬BC and ¬A¬BC potential instances to be added to the model. Since neither falsifies the model, we pick the one that minimizes negativity, in this case, A¬BC. Finally, the canonical model for the syllogism AI2 consists of the instances:

$$\begin{array}{ccc} \text{A} & \text{B} & \text{C} \\ \text{A} & \neg\text{B} & \text{C} \end{array}$$

Regarding the incorrect canonical and non-canonical models, we first derive all possible incorrect models for each syllogism. We then pick the one with the least amount of unique instances as the incorrect canonical model, and the one with the most as the incorrect non-canonical model. Similarly to above, if there are more than one possible choices, the one that minimizes negativity is chosen. For completeness, we also derive incorrect counterparts for the PMMs, by going through all possible incorrect models for a syllogism and employ a simple distance metric that measures the amount of different instances between two models. The most similar incorrect model is then chosen as an "incorrect PMM".

Every visual representation of a model consists of eight instances, since that is the maximum number of instances, should each one of them be different. If a derived model has less than eight instances, then some of them are repeated. In that case, we repeat the instances uniformly, while minimizing negativity, so that no bias is introduced because one instance appeared more than another. For the PMMs, we obtained the observed proportions by Todorovikj et al. (2023) and scaled them to our scenario of eight instances.

We found that three syllogisms have equal PMMs and canonical models (AA3, AI4 and EI3), so they have five corresponding tasks instead of six[2]. In total, we ultimately have $43 \times 6 + 3 \times 5 = 278$ tasks for 46 syllogisms. The participants are divided in five groups based on which syllogisms they are presented with. Following Todorovikj et al. (2023), we maintain a similar experience between participants by dividing the syllogisms in five groups based on their "preferedness", i.e. the construction frequency of the PMM. The final sets were then created by selecting one syllogism from each preference group, while ensuring that two syllogisms with a same quantifier order do not appear in the same set. That leads to four sets with nine syllogisms (two with 53 and two with 54 tasks) and one set with ten syllogisms and 59 tasks. The presented contents of the tasks were randomized per syllogism, per model. The resulting data and all materials are available on GitHub[3].

Table 2: Mean individual relative response time for each model and correctness. The full set consists of all tasks in the experiment, the reduced set eliminates the four syllogisms with less than 6 unique experimental tasks (AA3, AI4, EI3 and AA2).

Model	Full Set		Reduced Set	
	Correct	Incorrect	Correct	Incorrect
PMM	1.09	0.99	1.11	1.00
Canonical	1.03	0.80	1.04	0.82
Non-Canonical	1.19	0.89	1.19	0.90

[2] Due to a coding error in the experiment regarding the syllogisms AA2 and AA3, participants that answered for AA2 were not presented with its PMM and participants that answered for AA3 were presented with the same model twice (PMM = canonical). We include AA2's remaining responses when reporting response times and modeling, but not in the statistical analysis. For AA3 we only take into consideration the first appearance of the repeated task, to avoid potential, though unlikely, learning effects.

[3] https://github.com/saratdr/iccm-2024-SyllogisticPMMs

Participants

100 participants took part in our online web-experiment on the platform Prolific[4]. For the following analysis we performed a binomial test to determine an answer correctness percentage threshold (64-65%, depending on participant group, $p = .05$). We eliminated three participants whose correctness percentage was below the threshold, and two more due to technical issues. Ultimately, we have $N = 95$ participants (age 20-63, $M = 36.63$, $SD = 10.48$; 69% male). All of them were native English speakers. After completing the experiment, they received compensation of 6.75 GBP.

Procedure

At first participants are given an introductory task, where it's explained that they will be given two statements describing properties of shapes and are instructed to assume they are true. When they have read the statements, they are shown a visual representation of a set of shapes and are instructed that they will have to decide as quickly as possible if the set is in line with the statements or is contradicting them. Afterwards, the experiment starts, and participants are always presented with only the syllogistic premises at first. The experiment is self-paced, so once they decide to proceed, the visual representation of the model is shown as well, which they then have to accept that it corresponds to the premises or reject it. An example of a task is shown in Fig. 1.

Analysis

First, we analyzed to which extent the participants' responses were correct. Given a noteworthy correctness average of 91.61% with errors spread across all tasks, the verification task itself seemed to be so easy for all participants that errors can likely be accounted inattentiveness instead of a systematic mistake. Thus, we proceed with analyzing only correct answers. Note that because of that, throughout the analysis, the terms *correct* and *incorrect* always denote the properties of the respective model and do not refer to participants' response correctness. For our analysis, we rely on the response time between presentation of the model visualization and the participants' responses.

Table 3: Comparison of response times between correct and incorrect models in the reduced set using the Mann-Whitney U test. Significant p-values are marked in bold (corrected with Bonferroni-Holm method).

Model	Med. Corr.	Med. Incorr.	U	p
PMM	0.90	1.02	207562.5	<**.001**
Can	0.96	0.71	182432.5	<**.001**
NCan	1.05	0.79	180508.5	<**.001**

Annotation. Med. - Median; Corr. - Correct; Incorr. - Incorrect; Can - Canonical; NCan - Non-canonical.

Table 4: Comparison of response times between types of correct models in the reduced set using the Mann-Whitney U test. Significant p-values are marked in bold (corrected with Bonferroni-Holm method).

Models	Med. 1	Med. 2	U	p
PMM vs. Can	1.02	0.96	250408.0	**.034**
PMM vs. NCan	1.02	1.05	250798.5	**.034**
Can vs. NCan	0.96	1.05	237565.0	<**.001**

Annotation. Med. - Median; Can - Canonical; NCan - Non-canonical.

Since inter-individual differences can be substantial for response times and not necessarily reflecting the cognitive processes (i.e., the time needed to actually click on a response button), especially in online experiments, where the setup is non uniform, we standardized the recorded times for our analysis: For each task, we calculated the ratio between the respective response time and the overall mean response time of an individual. In the subsequent analysis we work with two sets of responses - the full set of all responses and a reduced set that does not contain responses for syllogisms with less than 6 unique tasks – AA3, AI4, EI3 and AA2. The first three have an equal PMM and canonical model, so a statistical comparison between those two models is generally impossible, whereas AA2 was affected by a coding error. Table 2 shows the mean individual relative response times for each correct and incorrect model, for both sets. Note that the impact of the elimination of the above mentioned syllogisms on the average times is negligible.

Focusing on the reduced set, we first examine the difference between correct and incorrect models. We can immediately notice that the incorrect canonical and non-canonical models were dismissed faster than the respective correct ones were accepted (0.82 vs. 1.04 for canonical; 0.90 vs. 1.19 for non-canonical). In the case of PMMs, though, there is a smaller difference (1.00 vs. 1.11), however, the increasing trend is still present. We tested for statistical significance in the changes using the Mann-Whitney U test and found that all differences are significant ($p < .001$), as shown in Table 3, along with the respective median values, for reference. This indicates that individuals are able to identify incorrect models faster than correct ones. This is plausible given that, for tasks with a universal quantifier involved (which are 40 out of the 46 tasks), participants can immediately reject the model once they recognize only one instance that contradicts the premises without even checking the rest, in contrast to correct models, where the whole model needs to be checked.

Next, we look into the response time differences between the three (correct) models. As intended, the canonical models represent the lower bound with 1.04 and the non-canonical ones the upper bound, with 1.19. The average response time for the PMMs lays in the middle with 1.11. Once again,

[4] https://www.prolific.co/

Table 5: Spearman correlation analysis between each type of instance and the mean response time for each model. Significant p-values are marked in bold (corrected with Bonferroni-Holm method). Note that instances that do not appear in a model or have a constant amount among all syllogisms lack a correlation value.

Inst.		PMM	Can	NCan	IPMM	ICan	INcan
Unique	M	3.25	2.56	5.93	2.90	1	7.72
	ρ	.15	.15	-.08	.03	–	-.26
	p	**.003**	**.001**	.559	1	–	<**.001**
Nec.	M	3.73	8	4.17	2.68	2.19	2.58
	ρ	-.05	–	.06	-.09	.06	.17
	p	1	–	1	.328	1	<**.001**
U. Nec.	M	1.57	2.56	2.56	0.86	0.27	2.37
	ρ	.07	.15	.04	-.07	.06	.01
	p	1	**.001**	1	1	1	1
Poss.	M	4.27	0	3.83	3.99	0.59	3.32
	ρ	.05	–	-.06	.18	.11	.18
	p	1	–	1	<**.001**	.075	<**.001**
U. Poss.	M	1.68	0	3.37	1.53	0.07	3.25
	ρ	.11	–	-.13	.15	.11	.18
	p	.056	–	**.013**	**.004**	.075	< **.001**
Inc.	M	0	0	0	1.34	5.22	2.10
	ρ	–	–	–	-.13	-.11	-.31
	p	–	–	–	**.024**	.056	<**.001**
U. Inc.	M	0	0	0	0.52	0.65	2.10
	ρ	–	–	–	-.10	-.11	-.31
	p	–	–	–	.192	.056	<**.001**

Annotation. Inst. - Instance (type); Can - Canonical (model); NCan - Non-canonical (model); IPMM/Ican/INcan - Incorrect versions of the models; U. - Unique; Nec. - Necessary; Poss. - Possible; Inc. - Incorrect.

we performed a Mann-Whitney U test and determined that the difference in response times between the models is statistically significant. Individuals needed more time to verify PMMs than canonical models ($p = .034$), but less than non-canonical ones ($p = .034$). Clearly, canonical models were evaluated faster than non-canonical ones ($p < .001$). All test results, along with the medians are displayed in Table 4. Note that all p-values reported in the two tables are corrected after the Bonferroni-Holm method for multiple comparisons. These results indicate that even though the PMMs are models that were the most frequently constructed ones, the time needed to verify such a model is not exceptionally short or long, falling between the two extreme bounds. In other words, in the domain of syllogistic reasoning, we cannot conclude that the preference for creating a model is related to the verification time or even its correctness, given the accuracy of above 90% across all models reported above.

Modeling

In this section we look into the structure of the given models, specifically, the type of their instances. We differentiate between: a) instances that are *necessary* for a correct model representation of a syllogism; b) instances that are *possible* to be added, i.e. do not contradict the premises, but aren't necessary and c) *incorrect* instances. Moreover we also look into *unique* instances, disregarding repetition. We analyzed the relationship between these descriptors and the mean relative response times. Table 5 shows the correlation results. We observe how different types of instances are significantly correlated with response times of different model, e.g. the canonical models correlate with the number of unique and unique necessary instances, whereas the non-canonical ones with the amount of unique possible instances. That is coherent with the definitions of the models relying heavily on necessary and possible instances, respectively.

As a next step, we investigate whether a model of the types of instances as descriptive features can successfully represent response times for each syllogistic model. To that end, we fit 127 linear regression models with ridge regularization, using all possible combinations of features of all lengths. We selected the best one based on the lowest Akaike Information Criterion (AIC; Akaike, 1974) and Bayesian Information Criterion (BIC; Schwarz, 1978) values. Using these two metrics on a full set of models, we can determine a threshold after which the addition of parameters does not lead to a significant fit improvement, while increasing the tendency of the models to overfit, and therefore select an appropriate model. Finally, we select the linear regression model considering the amount of unique necessary ($\beta = 0.06$), unique incorrect ($\beta = -0.11$) and possible ($\beta = 0.04$) instances, with $AIC = -854.71$ and $BIC = -843.86$, achieving a mean absolute error of $MAE = 0.16$. For a more detailed comparison, we reconstructed the mean relative response times based on the times predicted by the model, as displayed in Table 6. We observe results nearly matching the true values, while preserving the increase and decrease trends among different types of models and correctness. Thus, the model highlights how the time required by individuals for model verification is heavily based on its structure. This, again, corroborates our finding that potential preferences have little effect on participants' ability to verify models. Instead, the determining factor seems to be structural complexity of the model.

Table 6: Mean predicted relative response time for each model and correctness in the full set using a linear regression model with the number of *unique necessary*, *unique incorrect* and *possible* instances as features.

Model	Correct	Incorrect
PMM	1.11	1.00
Canonical	1.02	0.83
Non-canonical	1.16	0.91

Discussion

In this article, we continue the investigation of preferred mental models in the domain of syllogistic reasoning (Todorovikj et al., 2023) by posing two research questions. Analogous to PMM evaluation in the spatial domain (Ragni & Knauff, 2013), we first examine how trivial model verification is for individuals (**RQ1**). Thereby, we designed and conducted an experiment in the same world of marked, colourful shapes. This time, participants were presented with a syllogism and a corresponding model and asked to verify whether it is in line with the premises or contradicts them. We created six tasks per syllogism by deriving their canonical and non-canonical models as lower and upper bounds, respectively, using the already determined PMMs by Todorovikj et al. (2023), and deriving incorrect versions of all three model types, as control tasks. We found that individuals accept correct models and rejected incorrect ones with an average success of 91.61%, indicating that they do, in fact, verify models with ease, regardless of their structure.

For mReasoner (Khemlani & Johnson-Laird, 2013) and the Mental Model Theory in syllogistic reasoning, these findings have two implications: First, since models seem to be easily built and verified by human reasoners, the assumption that these processes do not involve errors is confirmed by our findings. Second, however, the fact that participants don't seem to need much effort for verification and construction, also raises the question, if the model manipulation during the search for counterexamples proposed by mReasoner is plausible: After all, an alternative solution could be to repeatedly rebuild different models instead.

Furthermore, we investigated whether a preference effect exists for accepting models in syllogistic reasoning (**RQ2**) by examining the response times for each model type and correctness. We found that the canonical and non-canonical models significantly represent the lower and upper bounds, as intended, while the mean response time for PMMs is in between them. Additionally, individuals needed significantly less time for rejecting incorrect models, respective to their counterparts, following the same trend of PMMs being in the middle of canonical and non-canonical models. This does not necessarily express any sort of *preference*, but seems to largely depend on the structural components of the models. Therefore, we analyzed the behavior further by describing the models using the types of instances they contain – necessary, possible, incorrect and unique – and finding significant correlations between them and the individuals' response times. Following that trace, we fit a set of 127 regression models, capturing all feature combinations, and found that the best representation uses the amount of unique necessary, unique incorrect and possible instances in a syllogistic model. Furthermore, the model was able to replicate the patterns in the data accurately, indicating that the selected structural properties are in fact sufficient.

In the empirical analysis of PMMs in spatial reasoning, Ragni and Knauff (2013) identified that the acceptance correctness of models constructed according to a preferred strategy is typically higher than for (correct) models built following a different one (92% vs. 81% and 44%). They report analog tendencies in the respective required response times as well (3.8*ms* vs. 4.36*ms* and 6.41*ms*). Similar findings are made by Rauh et al. (2005), who examined acceptance of conclusions following from the respective PMMs in spatial reasoning. Ultimately, we can conclude that individuals struggle with identifying and veryfing models that do not coincide with a preferred model/strategy in the spatial relational domain, but a similar conclusion can certainly not be made for syllogisms. In fact, we showed that the difference in required verification time depends on how "chaotic" a given model is and is not related to what was found to be preferred models. Logically, given a model with at least one instance contradicting the premises, the faster it's identified, the faster it will be rejected. The more frequent an instance is repeated in a model, the less time is necessary to verify all instances. Finally, a major difference between spatial reasoning tasks and syllogistic reasoning is in their typical experimental designs: During the whole duration of syllogistic reasoning tasks, both premises are usually visible, while they are only shown for a short duration (and one after the other) in many spatial reasoning experiments. It is plausible, that strategies allowing to quickly integrate new premises and without much load on working memory cause a preference for certain models to be built in spatial reasoning tasks, while the necessity is not present for typical syllogistic reasoning tasks.

So, what does this mean about preferred mental models in syllogistic reasoning? A few questions for future research and investigation arise: Why are most of the found PMMs not equal to the canonical models? It points to a tendency of individuals adding instances that are not directly observed in the premises, but also not to the extent that they reach a full fleshed-out non-canonical representation. There is a potential to interpret this as a way of communicating other possibilities exist and ensuring that this knowledge is accounted for. Though, is this a trend only among "simpler" syllogisms that by default do not require a large amount of necessary instances to represent them? Ultimately, an important point to consider is whether the reported preferred mental models are in fact the mental models individuals use to reason about a syllogism in the first place. Todorovikj et al. (2023) fit mReasoner to their data to show a lack of relevance of the mental models provided by the participants for the conclusions they provided later on. We can interpret the found preferred models as "prototypes" for a syllogistic model, however, cannot conclude that they are preferred models when reasoning, as it's done in the spatial domain.

Acknowledgements

This project has been partially funded by a grant to MR in the DFG-projects 529624975, 427257555 and 318378366.

References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE transactions on automatic control*, *19*(6), 716–723.

Johnson-Laird, P. N. (1975). Models of deduction. In R. J. Falmagne (Ed.), *Reasoning: Representation and process in children and adults* (pp. 7–54). New York, US: Psychology Press.

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

Johnson-Laird, P. N. (2008). *How we reason*. Oxford University Press.

Johnson-Laird, P. N. (2010). Mental models and human reasoning. In *National academy of sciences* (Vol. 107, pp. 18243–18250).

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., Lotstein, M., Trafton, J. G., & Johnson-Laird, P. N. (2015). Immediate inferences from quantified assertions. *The Quarterly Journal of Experimental Psychology*, *68*(10), 2073-2096. doi: 10.1080/17470218.2015.1007151

Knauff, M., Rauh, R., & Schlieder, C. (1995). Preferred mental models in qualitative spatial reasoning: A cognitive assessment of allen's calculus. In *Proceedings of the seventeenth annual conference of the cognitive science society* (pp. 200–205).

Ragni, M., & Knauff, M. (2013). A theory and a computational model of spatial reasoning with preferred mental models. *Psychological Review*, *120*(3), 561–588. doi: 10.1037/a0032460

Rauh, R., Hagen, C., Knauff, M., Kuss, T., Schlieder, C., & Strube, G. (2005). Preferred and Alternative Mental Models in Spatial Reasoning. *Spatial Cognition and Computation*, *5*, 239–269. doi: 10.1080/13875868.2005.9683805

Schwarz, G. (1978). Estimating the dimension of a model. *The annals of statistics*, 461–464.

Todorovikj, S., Brand, D., & Ragni, M. (2023). Preferred mental models in syllogistic reasoning. In *Proceedings of the 21st international conference on cognitive modeling.*