Do Models of Syllogistic Reasoning extend to Generalized Quantifiers?

Maximilian Mittenbühler (max.mittenbuhler@gmail.com)
Department of Computer Science, University of Freiburg, Germany

Daniel Brand (daniel.brand@hsw.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

Over the last century, a large variety of cognitive models for syllogistic reasoning have been developed, thereby advancing our understanding about the way humans process reasoning tasks. Most of the research was performed on a restricted set of quantifiers from first-order logic, which simplified model evaluations and comparison due to a well-defined set of tasks and the availability of complete and extensive datasets. However, as everyday reasoning and communication relies on a large variety of quantifiers, the scope and potentially also the generalizability of the models was severely limited. The present work aims at extending the domain of syllogistic reasoning to a wider set of quantifiers by (I) presenting a benchmarking dataset that includes the quantifiers "Most" and "Most not", (II) evaluating two state-of-the-art models (the Probability Heuristics Model and mReasoner) with respect to their ability to account for individual reasoners and (III) set the predictive performance of the cognitive models into perspective by comparing them to upper bounds and providing in-depth insights about their strengths and weaknesses.

Keywords: Syllogistic Reasoning; Generalized Quantifiers; Cognitive Modeling; Probability Heuristics Model; Mental Model Theory; mReasoner

Introduction

Syllogistic reasoning is one of the oldest domains for researching human reasoning capabilities, with a history of over a century (Störring, 1908). As an example, consider the following syllogism:

(1) Most Mammals are Land Creatures.
(2) Most Mammals are Intelligent Creatures.

What, if anything, follows from these two premises?

In general, syllogisms consist of two premises making a quantified statement about the relation between two terms (e.g., mammals and land creatures in the first premise), that are connected via a term occurring in both statements (middle-term; e.g., mammals). In this example, the task would be to infer the relation between the two end-terms (*land creatures* and *intelligent creatures*), which can be done by considering how each of them relates to *mammals*. In this case, it can be concluded that at least *some* land creatures are also intelligent (and therefore some intelligent creatures live on land). Generally, research has shown that humans systematically deviate from logic (e.g., Khemlani & Johnson-Laird, 2012), which prompted the development of theories that describe and explain how humans reason about such tasks.

Throughout the article, we will use common abbreviations (e.g., Pfeifer, 2006) for the syllogisms, using single letters for the quantifiers: *A*, *I*, *E*, *O*, *T* and *D* for *All*, *Some*, *No*, *Some...not*, *Most* and *Most...not*, respectively. Furthermore, we denote the order of the terms in the premises with a so-called *figure*. In this article, we use the definition of figures used by Khemlani & Johnson-Laird (2012), which is shown in the following table (leading to the abbreviation TT4 for the syllogism in the example):

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

Most research about syllogistic reasoning focused on a restricted subset of syllogisms that only considered the quantifiers from first-order logic (*All*, *Some*, *No* and *Some...not*, which we refer to as *classic quantifiers*) while excluding generalized quantifiers like *most* and *few*. This restriction has allowed researchers to investigate a well-defined subset of 64 possible syllogisms with 9 possible conclusions: 8 quantified conclusions (4 quantifiers with 2 directions each) and the option that there is no valid conclusion (NVC). Currently, a multitude of theories explaining how humans solve these syllogistic tasks exist (for an overview see Khemlani & Johnson-Laird, 2012), which were thoroughly evaluated in terms of their ability to predict general human behavior as well as adapt to individual reasoners (e.g., Khemlani & Johnson-Laird, 2012; Riesterer, Brand, & Ragni, 2020a). For these evaluations, *complete* datasets, i.e., where each participant solved all tasks of the domain (64 in this case), are pivotal as they allow an investigation on the level of individual participants without introducing a potential bias due to the task selection. Furthermore, purely data-driven models that require a rich data foundation can also be included as an upper bound for performance (Riesterer, Brand, & Ragni, 2020b).

Unfortunately, restricting the research focus to only four, first-order logic based quantifiers limits the applicability of the resulting theories to everyday communication and reasoning (e.g., Pfeifer, 2006), which involves a variety of qualitatively different quantifiers. The restriction therefore severely limits the scope of the understanding we obtained from our theories. However, while it would be beneficial to extend the set of quantifiers, it comes at a cost: Each additional quantifier exponentially increases the number of tasks, making the collection of a complete dataset challenging if not impossible. Selecting the quantifiers is also an arbitrary decision, as they are not part of an established framework, such as first-order logic that justifies the distinct restriction to a certain set.

To address this issue, we have collected a complete dataset with the additional quantifiers *Most* and *Most...not*, amounting to a total of 144 syllogisms per participant, in a recent study (Brand et al., in press). Importantly, these *generalized* quantifiers can not be expressed in first-order logic for sets of unknown sizes, which is usually the case for syllogistic tasks. Therefore, they could provide insight into a different facet of human syllogistic reasoning. Our analyses showed that the inclusion of additional quantifiers did not change the behavior on the classic syllogisms, leaving the validity of previous research efforts unchallenged. However, the vast majority of theories explaining syllogistic reasoning have exclusively been evaluated on the narrow set of classic syllogisms, and it remains unclear if these theories still apply to the wider domain of *generalized syllogisms*. To this end, the present work makes the following contributions: First, we repeated the study and collected additional participants in order to compile a dataset that is suitable for model evaluation and benchmarking in the domain of generalized syllogisms. Second, we evaluate two of the most prominent models for human syllogistic reasoning, mReasoner and the Probability Heuristics Model (PHM), which are both able to handle the additional quantifiers. We specifically focus on their capability to account for individual reasoning behavior as opposed to a distribution over a population. Finally, we analyze and discuss where the models succeed and where they fail at explaining human data by comparing them to several baseline models.

Related work

Probability Heuristics Model

The Probability Heuristics Model (PHM Chater & Oaksford, 1999; Oaksford & Chater, 2001) assumes that people's everyday reasoning does not follow logical validity of quantified assertions, but their probabilistic validity instead. The probabilistic validity (or p-validity) of a conclusion is defined by the conditional probability of the end-terms, which in term is determined by the conditional probabilities described in the premises (where an end-term is one of the terms that are to be connected in the syllogistic task). The PHM proposes that people do not deduce p-validity mathematically but instead use a number of heuristics that converge to p-validity. These heuristics are based on the notion of p-entailment, describing that certain quantifiers probabilistically follow from others (for instance, ”All” entails ”Some”), and the notion of informativeness, detailing that less probable and therefore more specific quantified assertions are more informative. This yields the informativeness order of quantifiers: *All* $>$ *Most* $>$ *Most not* $>$ *Some* $>$ *No* $\geq$ *Some not*. To generate a conclusion candidate, the PHM uses the following three generative heuristics (G1-G3): First, the *min-heuristic* (G1) identifies the premise with minimal informativeness (min-premise) to determine the quantifier of the conclusion. Second, an alternative candidate quantifier that probabilistically follows the quantifier from G1 is proposed (*p-entailment*, G2). Finally, the direction of the conclusion is determined by the *attachement heuristic* (G3). If the min-premise from G1 starts with an end-term, the respective term is used as the subject of the conclusion. Otherwise, the end-term of the remaining premise (max-premise) that features the most informative quantifier is used as the subject of the conclusion.

The PHM also assumes that people may test their initial deductions. It proposes that this process comprises a further two heuristics (T1 and T2), which evaluate how much confidence should be granted to the conclusion candidate (either the candidate with the quantifier determined by G1 or G2). To this end, the informativeness of the max-premise is considered by the *max-heuristic* (T1). It is assumed that confidence and the informativeness of the max-premise are coupled, which means that NVC can be concluded if the confidence is too low (Copeland, 2006). Additionally, the *O-heuristic* (T2) postulates that *Some not* (O) should generally be avoided in conclusions due to their lack of informativeness. However, given the mechanism of the max-heuristic, O-conclusions already are the conclusions with the lowest confidence, which makes the O-heuristic more a refinement than an independent heuristic.

It is important to note that the interpretation of the quantifiers assumed by PHM excludes *All* from the quantifier *Most* (i.e., if *Most A are B*, then *All A are B* does not hold). However, *Some* also includes the possibility of *All*, following the traditional interpretation from first-order logic. Negated quantifiers are treated analogously.

mReasoner

Another prominent theory for syllogistic reasoning is the Mental Model Theory (MMT; e.g., Johnson-Laird, 2010). MMT assumes that reasoners construct a mental model representing the information provided by the premises of the syllogism that is then used to derive a conclusion. It thereby follows a four-step procedure (Copeland, 2006): The first premise is used to create a mental model representing the information by an instantiated set of entities that are assigned to the syllogistic terms of the premise based on the respective quantifier. Then, the mental model is extended by the second premise, thereby integrating information about the third syllogistic term. In the third step, a conclusion candidate is derived from the mental model. Finally, the conclusion candidate is tested by a search for counterexamples, that checks if the conclusion candidate holds up to alternative mental models that are consistent with the premises. If a counterexample is found, the mental model is either corrected and a new conclusion candidate is derived, or the process is aborted and NVC is concluded. If no counterexample is found, the candidate is accepted as the conclusion to the syllogism.

This process is implemented in the LISP-based cognitive model mReasoner[1] (Khemlani & Johnson-Laird, 2013). It uses four parameters associated with the inference process (Khemlani & Johnson-Laird, 2016): λ determines the maximum number of entities in the initial mental model by specifying a Poisson distribution from which the number of entities is drawn. ε then determines the completeness at which the premise information is represented within the entities. Finally, σ controls the likelihood to engage in the search for counterexamples. ω then controls the behavior of mReasoner in the case that a counterexample was found by specifying the probability of weakening the conclusion quantifier and re-engaging in the search for counterexamples. If

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

a counterexample was found and the conclusion quantifier is not weakened, NVC is concluded instead.

Expanding mReasoner to generalized quantifiers Building mental models of quantified assertions containing generalized quantifiers poses a particular challenge to mReasoner because of the ambiguity of the quantifiers *most* and *most not* under certain circumstances (S. Khemlani, personal communication, March 3, 2022). To incorporate this, the Authors have equipped mReasoner with a more general model-building system than that required for syllogisms that only contain the classic quantifiers. More specifically, it takes advantage of its ability to generate mental models of different sizes (governed by its λ-parameter) as well as its stochastic mode. By incorporating the ability to parse generalized quantifiers in the stochastic model-building system, mReasoner can represent statements containing ”most” in figure 2 or 3, which it would not be able to do otherwise (S. Khemlani, personal communication, March 3, 2022).

Method

Data

In a previous study, the responses of 31 participants to 144 syllogisms were collected over the course of three sessions in order to minimize fatigue (Brand et al., in press). The study comprised all 64 syllogisms with the first-order logic quantifiers *All*, *Some*, *No* and *Some not* as well as 80 additional tasks consisting of syllogisms with the generalized quantifiers *Most* and *Most not*. To minimize biases due to the content of the syllogisms, hobbies and professions were used for the terms. The study thereby covered all syllogisms that could be constructed from the 6 quantifiers. Participants were asked to give either a quantified conclusion following from the premises or to respond with *No valid conclusion*, if no conclusion was possible. For the present work, we re-ran the study and extended the dataset by another 34 participants. The following analysis is therefore performed on a dataset consisting of 65 participants (mean age: 39.1, age SD: 14.0, female: 52.3%), where each responded to all 144 syllogistic tasks. The dataset and materials for the analysis are publicly available on GitHub[2]. Note that for assessing the correctness of participants' responses, we use the common interpretation that $Most(A,B)$ for finite sets A and B as $|A \cap B| > |A - B|$, with $|\cdot|$ being the size or the number of their elements (e.g., Westerståhl, 1989; Novák, 2008). Therefore, we are treating *Most* as *More than half*, which means that *All* also implies *Most*. However, no specific interpretation for the quantifiers was instructed in the study, so that the participants' understanding of the quantifiers are reflected in their response behavior.

Model Evaluation

For the following analyses, we used the Cognitive Computation for Behavioral Reasoning Analysis (CCOBRA) framework[3] and its *coverage* evaluation type (see Riesterer, Brand, & Ragni, 2020a). In this type of evaluation, the parameters of both PHM and mReasoner are first optimized for each participant by grid searching the parameter space and selecting those parameter settings that yield optimal mean accuracy. Using the optimal parameter settings for each participant, the models are then queried for predictions of the responses that the participant gave for all tasks. Overall model predictive performance is assessed via the achieved accuracy. Technically, the models were thereby fitted to the exact responses that it later has to predict. This means that a fully data-driven model with no restrictions on the number of parameters would be able to achieve a perfect prediction. However, cognitive models are restricted by the number and expressiveness of their parameters: The parameters should reflect and control meaningful mechanisms in the model's processes. Therefore, the *coverage* evaluation assesses the models' capabilities to represent the individual response patterns within the framework of their assumed processes and mechanisms and by that explaining the individual behavior.

PHM

In the following analyses we build upon a recent Python-based implementation of PHM, which used binary parameters to fit the model to individual reasoners (Riesterer, Brand, & Ragni, 2020a). In their implementation, a parameter for each confidence in a certain quantifier was implemented. Additionally, a parameter was introduced for the p-entailment, which specified if the conclusion based on the min-heuristic or the p-entailment should be used. While the parameters are usually continuous and interpreted as probabilities, the implementation was aiming at individual reasoners instead of a group of reasoners. Therefore, the parameters could be binary: As each participant usually only solves each task once, a prediction of the specific response has to be achieved by a model, instead of a distribution of possible responses. This simplifies the fitting process, as the number of parameters is quite low and allows for a exhaustive grid search in the parameter space. Additionally, the parameter space is further restricted by the additional constraint that the confidences follow the same ordering as the informativeness. Therefore, the confidence for *Some* can never be higher than the confidence for *All*. As the original implementation by Riesterer, Brand, & Ragni (2020a) only considered the 4 quantifiers from first-order logic, we extended the model to the generalized quantifiers *Most* and *Most not*. It is important to note that we incorporated *Most not* in the same way as *Few* was used in the original description of PHM by Chater & Oaksford (1999).

Furthermore, Chater & Oaksford (1999) also consider *weak p-entailment*, which would allow *Most* and *Most not* to follow from the quantifiers *Some* and *Some not*. In our implementation, we do not consider weak p-entailment, which implies that generalized quantifiers in conclusions are never considered for the classic syllogisms.

mReasoner

For mReasoner, we used the Python-based model by Riesterer, Brand, & Ragni (2020a) which internally relies on the original LISP-implementation of mReasoner in order to rule out differences in the model behavior. The model was then extended to the quantifiers *Most* and *Most not*, and the updated version of mReasoner was used. The parameters were fitted using a grid-seach with 6 steps for each parameter. For ε, ω and σ which have a range from 0 to 1, this yiels a stepsize of 0.2. The range

[2] https://github.com/Shadownox/iccm-syl-genquant-models

[3] https://github.com/CognitiveComputationLab/ccobra

for λ was chosen to be between 3 and 8 (which leads to a stepsize of 1). While Riesterer, Brand, & Ragni (2020a) used the full range of λ with $\lambda \in [0,8]$, the extension to generalized quantifiers required higher values to work. Furthermore, it was required that $\varepsilon < 1$. To account for the randomized nature of the inference process, each configuration was sampled 10 times.

Baseline Models

Similar to existing benchmarking settings for syllogistic reasoning (e.g., Brand et al., 2020; Riesterer, Brand, & Ragni, 2020a,b), we included a *Random* model as a lower bound of the performance, which uniformely selects one of the possible response options, as well as the most-frequent answer (*MFA*), which uses the most frequently given response to a syllogism as a prediction. The MFA is also the best model when not fitting to individual participants. To assess the maximum predictive performance (theoretically) achievable with the present dataset, we included a purely data-driven model as an upper bound (for a similar application of data-driven models, see Riesterer, Brand, & Ragni, 2020b). We used a user-based collaborative filtering model (*UBCF*), which is a neighborhood-based model from the field of recommender systems that relies on the behavior of other *users* to predict a targets' behavior (for an in-depth description, see Aggarwal, 2016). Based on the responses given to all syllogisms except for the one to be predicted, a neighborhood of the k most similar participants is created. When predicting the response of a target participant to a syllogism, each neighbor *votes* for the responses, where the vote is weighted by the respective similarity to the target participant. To discount less similar neighbors even more, the similarity can be raised to the power of an exponent-factor *exp*. The final prediction is then the response with the most votes. For this analysis, we used the parameters $k = 12$ and $exp = 3$, which was found by applying a grid-search for the best parameters. One advantage of the UBCF is the similarity to the MFA, as the MFA can be interpreted as a special case of the UBCF: If no information about the target participant is available, the similarity is not defined, leading to the neighborhood consisting of all other participants available. Therefore, the prediction would just be the most frequently given response. Therefore, the UBCF can be considered as an extension of the MFA to the individual level.

Analysis

Overall Model Performance

Figure 1 shows how well mReasoner and PHM, as well as the three baseline models, were able to predict participants' responses. Both mReasoner, with on average 39.7% correct predictions, and PHM, with 41.7% correct predictions, performed noticeably above chance-level at 7.7% and were able to surpass the MFA-model at 35.6%. The general performance indicates that both models can at least partly explain peoples' responses. The difference to the MFA-model did, however, not reach significance (Mann-Whitney-U test: $U = 1882.5$, $p = 0.29$ for mReasoner, $U = 1780.5$, $p = 0.12$ for PHM, respectively), which shows that the ability to adjust to individual response behavior is still lacking, which is also corroborated by the performance of the UBCF model with 45.2%. It becomes apparent that there is still a substantial amount of information available in the data, which is not yet covered by the models' mechanisms. Despite the general problems with adapting to individual reasoners, both cognitive models seem to be able to adapt to a small group of reasoners exceptionally well, indicating that the models generally are able to adapt to individuals, but still miss out on important mechanisms. This highlights the potential for further improvements of cognitive models for syllogistic reasoning.

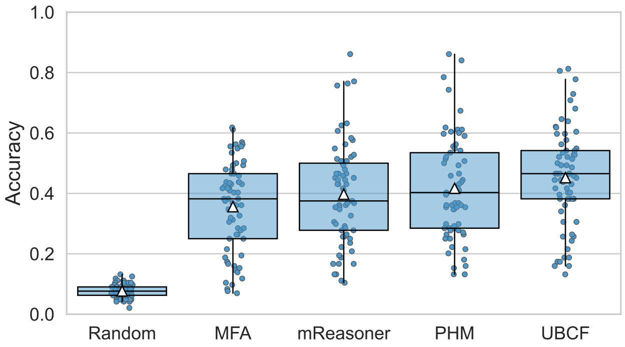


Figure 1: Predictive accuracy of the tested models on all syllogisms. Each point represents the accuracy for predicting a specific participant. The triangle denotes the respective mean.

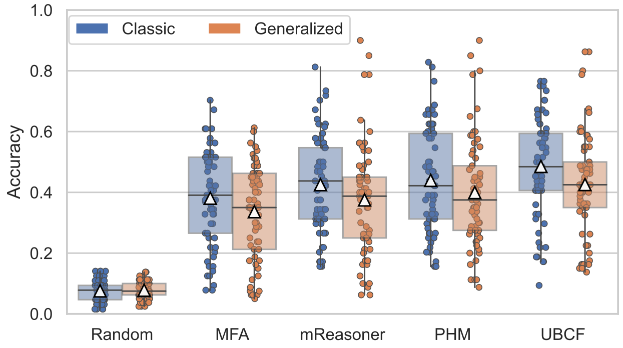
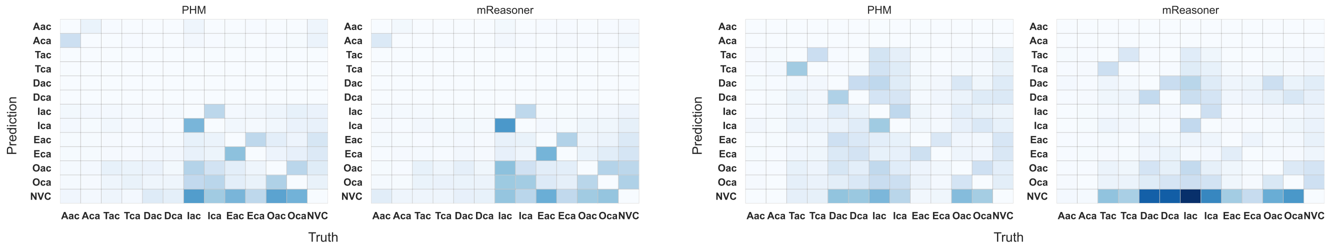


Figure 2: Predictive accuracy of the tested models for classic syllogisms (blue) and generalized syllogisms (orange). Each point represents the accuracy for predicting a specific participant. Triangles depict mean accuracy scores.

Performance for Classic and Generalized Quantifiers

As our focus was on expanding mReasoner and PHM to the domain of generalized quantified syllogisms, the differences in the model performance between the two domains are especially important. Therefore, Figure 2 depicts the results broken down by the respective task domain (i.e., classic syllogisms and syllogisms with generalized quantifiers). Note that, like in the general performance analysis, the models are still fitted based on all tasks, as we aim at evaluating the models' abilities to generalize across the different task types. It becomes apparent that all models perform worse on generalized quantified assertions by about five percentage points (except for the chance-level baseline).



(a) Errors on classic syllogisms

(b) Errors on syllogisms with generalized quantifiers

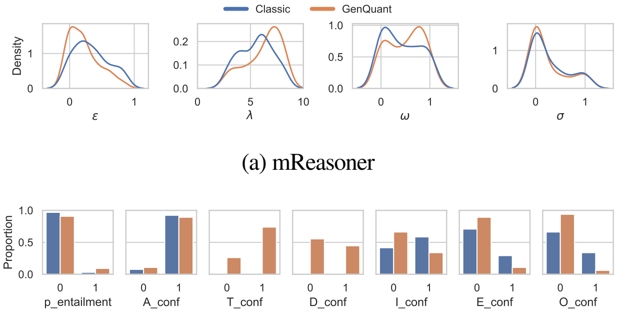
Figure 3: Errors when predicting responses for classic syllogisms (a) and syllogisms with generalized quantifiers (b) for PHM and mReasoner, respectively. Darker blue colors denote a higher number of errors. Both models were fitted on the full set of syllogisms.

However, the fact that the UBCF model's performance dropped to a similar extent indicates that this drop could be attributed to the participants' response behavior being less clear. This is corroborated by the fact that classic syllogisms were easier for the participants to solve (mean correctness: *GenQuant* = 0.25; *Classic* = 0.34), which in turn can minimize individual differences for some tasks (i.e., if there is an obvious answer). Yet again, a wider range of responses to generalized syllogisms could not be found: We compared the entropy (see Shannon, 1948) as a metric for uncertainty of the participants' response distributions for both, the classic and the generalized quantifiers, in order to check for a systematic difference in the range of responses. The entropies showed no substantial difference between both task types (*GenQuant* = 3.30; *Classic* = 3.22). However, easier tasks can nevertheless help to improve the consistency within participants' responses (i.e., the participant would reliably show the same response patterns), which makes it easier for models to replicate the response pattern, which might explain the differences between both task types.

Error Analysis

To see where the predictions of the cognitive models did not capture the human responses well, we investigated for which responses the most errors occurred (see Figure 3). For PHM, an indistinct picture emerges. While it seems that PHM generally tends to respond NVC too frequently, it does so for both task types in a comparable fashion. It also seems to misjudge the direction of the conclusion when not responding with NVC in both task types. However, while the errors based on NVC and the direction explain the majority of the errors on the classic syllogisms (65,4%), this does not hold for the generalized quantifiers (49%): Here, PHM also often mixes the quantifiers up, especially between I, D and O. It seems to be the case that participants are more variable in their use of these quantifiers as to the fixed order of informativeness PHM relies on.

When focusing on the results for mReasoner, a much clearer picture emerges. While the errors on the classic syllogisms are rather similar to the errors shown by PHM, NVC accounts for the vast majority of errors for the generalized quantifiers. NVC is the logically correct response for the majority of tasks, especially for generalized syllogisms (*GenQuant* = 76.3%, *Classic* = 57.8%), which seems to be reflected in mReasoner's mechanisms. However, this is not reflected in the participants responses, which do not show a difference in their NVC response behavior (*GenQuant* = 21.2%, *Classic* = 21.6%). Furthermore, mReasoner's mechanisms for giving NVC-responses seem to be too coarse: If it needs to respond with NVC for several tasks, it seems to overshoot substantially. The differences between classic and generalized syllogisms also seem to reflect that mReasoner handles generalized quantifiers differently than the classic quantifiers.



(a) mReasoner

(b) PHM

Figure 4: Parameter distributions after fitting mReasoner and PHM to each participant for the classic syllogims (blue) and syllogisms with generalized quantifiers (orange).

Parameter Analysis

Based on previous analysis, we investigated the parameters that the models would use for both task domains when fitted to them separately. Figure 4 shows the parameter distributions for both models when fitted to the responses of each individual participant on the classic syllogisms and the generalized syllogisms, respectively. Interestingly, the parameters of mReasoner do not show substantial differences except for ω, which controls behavior when a counterexample is found. While mReasoner was shown to respond with NVC too frequently, the difference in ω indicates that NVC was in fact moderated by the parameters, as it means that a conclusion in case of a found counterexample is rather weakened than directly concluding NVC. Generally though, the parameters indicate that the mReasoner's performance would not change much if fitted to the generalized quantifiers directly, which implies that the performance was not impeded by a generalizability problem (i.e., having to find parameters that work for both, classic and generalized syllogisms), but rather due to a general inability to account for certain response patterns occurring for generalized syllogisms. For PHM, the results are generally more shifted towards responding with NVC for the generalized quantifiers, by having a lower confidence for all quantifiers. Although differences between both task types show, the adaption to generalized tasks is mainly done by the specific parameters for the quantifiers T and D, which do not affect the classic tasks, as T and D can only become conclusion candidates if they are present in the premises (note that this would change if weak p-entailment was considered). In this regard, PHM has a distinct advantage over mReasoner, as it utilizes parameters that are specific for the extension to generalized quantifiers, while mReasoner relies on the same core paramters for all tasks.

Discussion

In this work, we performed a thorough evaluation of the predictive capabilities of PHM and mReasoner when confronted with syllogistic reasoning tasks that include the generalized quantifiers *Most* and *Most not*. The evaluation was performed on a benchmarking dataset that contains the responses to all 144 syllogisms for all participants, which allowed an analysis on the level of individual participants. The cognitive models were compared with the most-frequent answer and an estimated upper-bound given by a data-driven model based on user-based collaborative filtering. Both cognitive models performed within expectations, as they managed to slightly surpass the MFA, although not significantly. However, a more detailed look into the performance for individual participants, it appears that they are able to capture some of the participants well and seem generally able to adapt to individual participants. However, their performance fell short of to the UBCF, which highlights the potential that is still left in the domain and indicates that the models' mechanisms are still not sufficient to cover the variety of response patterns shown by different individuals.

When focusing on the generalized quantifiers, the performance of all models dropped substantially (including the UBCF), which indicates that the noise-levels are higher on these tasks. This is supported by the lower correctness on these tasks, which can lead to less consistent response behavior. However, the cognitive models still managed to surpass the performance of the MFA, which shows that their general mechanisms can generalize from the four first-order logic quantifiers to an extended set of quantifiers. This is corroborated by an analysis of their parameters, which showed no substantial differences when fitted to the classic tasks or the generalized tasks only.

Given the performance of both models, no difference, on neither the classic nor the generalized syllogisms, is noticeable. Therefore, based on the predictive performance, the assumed underlying processes both seem to be equally plausible. However, when the errors of both models are analyzed in detail, differences become apparent. As it was already shown that models have difficulties with correctly predicting the NVC-response on the classic syllogisms (Riesterer, Brand, Dames, & Ragni, 2020), it was likely that the problem carried over and thereby accounted for a part of the errors. This shows for both models across both task types, with NVC being an important source of error. However, the magnitude of the problem greatly differs between the models: On the one hand, the type of errors of PHM remain largely the comparable between classic and generalized syllogisms with NVC- and direction-related errors, despite an increase in noise-like errors on the generalized tasks. On the other hand, mReasoner fails to replicate the participants' NVC-behavior and drastically overshoots with the frequency of NVC responses on generalized syllogisms, while being comparable to PHM on classic tasks. This indicates that its mechanism for handling generalized syllogisms is currently inferior to PHM, although the problem seems to be covered by the high number of NVC responses that make predicting NVC frequently a rather safe strategy.

However, even though mReasoner currently seems to lag a bit behind, it is important to note that PHM utilizes specific parameters for the respective quantifiers, while mReasoner relies on a fixed set of parameters and its core mechanisms. This can greatly affect the future development, as it will be important to further extend the scope of the domain in order to advance our understanding in the field of syllogistic reasoning. While PHM can be rather easily adapted to additional quantifiers, it also means that the complexity of the model increases directly with the number of supported quantifiers, which can become an important factor when extending the domain further.

By providing a complete dataset and an evaluation of two state-of-the-art models, the present work aims at setting a starting point for extending modeling endeavors to an extended set of syllogisms. However, a large variety of other quantifiers are important for our everyday reasoning and communication, including more vague quantifiers like *Many* or counting quantifiers (e.g., *More than 3*). These possibilities have to be investigated in the context of syllogistic reasoning, in order to warrant the claim that the present models and our knowledge reaches beyond well-defined abstract tasks.

Acknowledgements

Support to MR by the German Research Foundation, DFG (RA 1934/4-1, RA1934/8-1) is gratefully acknowledged.

References

Aggarwal, C. C. (2016). Neighborhood-based collaborative filtering. In *Recommender systems: The textbook* (pp. 29–70). Cham: Springer International Publishing.

Brand, D., Mittenbühler, M., & Ragni, M. (in press). Generalizing syllogistic reasoning: Extending syllogisms to general quantifiers. In *Proceedings of the 44th Annual Conference of the Cognitive Science Society.*

Brand, D., Riesterer, N., & Ragni, M. (2020). Extending TransSet: An individualized model for human syllogistic reasoning. In T. C. Stewart (Ed.), *Proceedings of the 18th International Conference on Cognitive Modeling* (pp. 17–22). University Park, PA: Applied Cognitive Science Lab, Penn State.

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

Copeland, D. E. (2006). Theories of categorical reasoning and extended syllogisms. *Thinking & Reasoning*, *12*(4), 379–412.

Johnson-Laird, P. N. (2010). Mental models and human reasoning. *Proceedings of the National Academy of Sciences*, *107*(43), 18243–18250.

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.

Khemlani, S. S., & Johnson-Laird, P. N. (2016). How people differ in syllogistic reasoning. In A. Papafragou, D. Grodner, D. Mirman, & J. C. Trueswell (Eds.), *Proceedings of the 38th Annual Conference of the Cognitive Science Society* (pp. 2165–2170). Cognitive Science Society.

Novák, V. (2008). A formal theory of intermediate quantifiers. *Fuzzy Sets Syst.*, *159*, 1229-1246.

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

Pfeifer, N. (2006). Contemporary syllogistics: Comparative and quantitative syllogisms. In G. Kreuzbauer & G. J. W. Dorn (Eds.), *Argumentation in Theorie und Praxis: Philosophie und Didaktik des Argumentierens* (pp. 57–71). Wien: LIT.

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. (2020a). 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.

Riesterer, N., Brand, D., & Ragni, M. (2020b). Predictive modeling of individual human cognition: Upper bounds and a new perspective on performance. *Topics in Cognitive Science*, *12*(3), 960–974.

Shannon, C. E. (1948). A mathematical theory of communication. *The Bell System Technical Journal*, *27*(3), 379-423.

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

Westerståhl, D. (1989). Aristotelian syllogisms and generalized quantifiers. *Studia Logica*, *48*(4), 577–585. doi: 10.1007/BF00370209