Uncovering Iconic Patterns of Syllogistic Reasoning: A Clustering Analysis

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

Nicolas Riesterer (riestern@cs.uni-freiburg.de)
Cognitive Computation Lab, University of Freiburg, Germany

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

Abstract

Syllogistic reasoning is one of the core domains of human reasoning research. Over its century of being actively researched, various theories have been proposed attempting to disentangle and explain the various strategies human reasoners are relying on. In this article we propose a data-driven approach to behaviorally cluster reasoners into archetypal groups based on non-negative matrix factorization. The identified clusters are interpreted in the context of state-of-the-art theories in the field and analyzed based on the posited key assumptions, e.g., the dual-processing account. We show interesting contradictions that add to a growing body of evidence suggesting shortcomings of the current state of the art in syllogistic reasoning research and discuss possibilities of overcoming them.

Keywords: syllogistic reasoning; cognitive modeling; clustering; non-negative matrix factorization; dual-process theory

Introduction

The ability to reason about information is an essential skill for humans in almost all aspects of their lives. Consequently, the research of human reasoning has been a key field of study to advance our understanding about human cognition for an extensive time span. One of the core domains within the field is syllogistic reasoning, which is being investigated for over a century now (Störring, 1908). In its most common form, a syllogism consist of two quantified statements (premises) with first-order logic quantifiers (*All*, *Some*, *No*, and *Some ... not*), that interrelate three terms (commonly abbreviated by A, B, and C) via a middle-term as shown in the following example:

All A are B.
Some B are C.
What, if anything, follows?

The task is to conclude what the relation between the two end-terms occurring in only one of the premises (A and C) is. Additionally, there is the possibility that no valid conclusion (NVC) is possible resulting in a total of nine distinct response options.

For convenience, syllogisms are often abbreviated. Quantifiers are represented by the letters (*All*: A, *No*: E, *Some*: I and *Some ... not*: O). The arrangement of the terms is called *figure*. Throughout this paper, we will use the figure notation from Khemlani and Johnson-Laird (2012), which is shown in the table below:

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

Put together, the syllogism in the example above would be abbreviated as *AI1*. Conclusions can be represented in a similar way, combining the quantifier and the direction (*ac* or *ca*). For example, *Some C are A* would be abbreviated by *Ica*.

Given the long history of research, it is not surprising that a large variety of competing theories and models exist. However, the field was unable to reach consensus: In a recent meta-analysis, twelve theories of syllogistic reasoning were compiled and evaluated, concluding that "none of the existing theories is correct. Investigators of reasoning need to develop a better theory of monadic reasoning." (Khemlani & Johnson-Laird, 2012, p. 23).

Since the inferential mechanisms and strategies are substantially influenced by individual factors (e.g., working memory Gilhooly, Logie, Wetherick, & Wynn, 1993) and are susceptible to the influence of external factors (e.g., content and personal beliefs; Morgan & Morton, 1944), it is not surprising that the observed reasoning behavior shows significant inter-individual differences (e.g., Dames, Klauer, & Ragni, 2022) that current models struggle to capture (Riesterer, Brand, & Ragni, 2020a). As human reasoning behavior seems to be highly individual, the idea that no single inferential account may be able to capture every individual suggests itself (Khemlani & Johnson-Laird, 2012). Consequently, it seems to be more sensible to try to disentangle the different reasoning strategies.

From a more abstract data-driven perspective, the observed behavior of a human reasoner can be represented as a task-response-pattern. A model accounting for the behavior then specifies a process that generates the respective pattern. Thereby, it is restricted by its assumptions and the corresponding parameter space (for a model-evaluation based on this principle, see Riesterer et al., 2020a). From this perspective, the question of disentangling different strategies can be reformulated as a question of uncovering a set of latent (iconic) patterns that are suitable for capturing the patterns of most individuals to a satisfying degree. In this work, we utilize clustering methods to uncover the latent response patterns of individual reasoners and present a way to determine the number of central reasoning strategies.

The rest of the article is structured as follows: First, the background relevant to this work will be introduced. Second, our dataset and the clustering approach used to extract the iconic patterns are described. Third, the obtained patterns are interpreted with respect to their meaning for the state of the art in syllogistic reasoning. Finally, the results are discussed and a general outlook is given.

Background

A common approach aiming at describing the behavioral differences observed in reasoning is the dual-processing account (Evans, 2008), which proposes two mechanisms: a fast-and-frugal heuristic approach (*System 1*; S1) and a deliberative, more logical mechanism (*System 2*; S2). In the field of syllogistic reasoning, models often fall clearly into one of the two categories, with heuristics (e.g., PHM; Chater & Oaksford, 1999) belonging to S1 while approaches closer to logic (e.g., PSYCOP; Rips, 1994) would generally be considered to rely on S2.

Probably the most prominent theory incorporating the idea of dual-processing is the Mental Model Theory (MMT; Johnson-Laird, 1975) and its implementation mReasoner (Khemlani & Johnson-Laird, 2013). At its core, MMT assumes that syllogistic inference is a three-step procedure (Bara, Bucciarelli, & Johnson-Laird, 1995). In the first step, the premises are interpreted to construct a mental model representation of the information. This model is then extended to also incorporate the information of the second premise. In the second step, the constructed model is used to derive a conclusion candidate. This candidate is then put to the test in the third step, which consists of a search for counterexamples, i.e., models that contradict the conclusion but are still consistent with the premise information. If no counterexample is found, the candidate will be responded as the conclusion. Otherwise, a new conclusion candidate is generated, which is then subjected to the search for counterexamples again, or it is concluded that “no valid conclusion” is possible if no new candidates can be created.

The expensive search for counterexamples in MMT is assumed to be a S2 process, while conclusions directly inferred from the initial mental model reflect the more intuition-based strategy associated with S1.

It is important to note that while the average correctness of a participant’s responses typically increase with a higher number of NVC responses (Dames et al., 2022), seemingly corroborating the notion of S2 being responsible for NVC responses, invalid syllogisms are over-represented in the syllogistic domain with more than half of the syllogisms being invalid despite NVC being only one out of nine possible responses. Furthermore, recent work found that the response times did not increase for NVC responses as it would be assumed when engaging in a exhaustive search for counterexamples (Brand, Riesterer, & Ragni, 2022), sowing doubt if the proposed distinction into S1 and S2 truly reflects the processes underlying syllogistic reasoning.

Method

Dataset

The foundation of our analysis is a publicly available dataset by Dames et al. (2022), which contains the response data of 106 participants to all 64 syllogistic tasks. In the original analysis, participants were asked to complete all 64 tasks twice to investigate potential retest effects. However, these effects are out of scope for the present work and the respective data from the syllogistic retest is therefore excluded. Additionally, a variety of individual information about the participants is provided, out of which the *Cognitive Reflection Test* (CRT; Frederick, 2005) including additional questions by Toplak, West, and Stanovich (2014) and the participants’ Need for Cognition (NFC; see Cacioppo & Petty, 1982) are relevant for this work.

Clustering

Clustering refers to an unsupervised learning process of grouping objects together that are similar with respect to some similarity measure (for an overview, see Aggarwal, 2015). The clustering methods used in this work are thereby partitional approaches that grouping objects into disjoint sets by minimizing a cost function (e.g., euclidean distances between objects and cluster centroids in k-Means clustering). For our analysis, we compare the performance of k-Means, k-Medoids and a clustering method based on Non-Negative Matrix Factorization (for a similar method, see J. Kim & Park, 2008). As k-Means and k-Medoids are standard procedures, they will only briefly be discussed with respect to potential strengths and weaknesses for the specific analysis.

Of the three methods, k-Means is probably of the most prominent approach for cluster analyses. As the name suggests, k-Means divides objects into k clusters that are defined by centroids representing the mean of the respective objects in the cluster. Thereby, it behaves similar to an aggregation of the data that is commonly performed to investigate response distributions, with the difference that k distributions are obtained instead of a single one, thereby having the potential to provide a better fit for individuals. However, aggregation of data has been criticized to be problematic when investigating individual processes (Riesterer, Brand, & Ragni, 2020c), as different strategies might be entangled by the aggregation process.

In contrast to k-Means, k-Medoids uses actual datapoints as the centroids of the clusters. Hence, no aggregation is performed, eliminating the problems associated with it. However, as the number of participants in reasoning experiments is very limited compared to typical datasets used in machine learning, the approach might not find an optimal centroid for each cluster. Since human data is inherently prone to noise, a pattern found by k-Medoids might contain artefacts that were introduced by confounders unrelated to reasoning processes.

Clustering using NMF Non-Negative Matrix Factorization has the goal of finding a decomposition for an input-matrix X. To this end, a basis matrix $W = m \times k$ and a coefficient matrix $H = n \times k$ for a given k need to be found such that:

$$X \approx WH^T \tag{1}$$

These matrices can be obtained by using a variety of solvers, including the commonly used non-negative least squares solver (H. Kim & Park, 2008).

Formally, clustering can also be understood as a problem of matrix decomposition (e.g., J. Kim & Park, 2008). The columns in the W-matrix then represent the centroids of a cluster, while the H-matrix contains the assignment of a data point to the respective cluster.

To use NMF clustering on the syllogistic data, it needs to be represented as a matrix X of shape $m \times n$, where n corresponds to the number of participants and each column corresponds to an m-dimensional vector representing the respective participant's response pattern. To transform the data accordingly, we first represented the responses of each participant as a 64×9 matrix (for the 9 possible response options), meaning that each task is represented as a one-hot-encoded vector. The matrices were subsequently flattened into a 576-dimensional vector, out of which the final data matrix X containing all participant vectors was created (leading to $X = 576 \times 106$). In order to find the matrices W and H, we used the non-negative least squares solver included in the Python package SciPy[1] which is based on the algorithm proposed by Lawson and Hanson (1995).

In order to realize clustering via NMF, additional constraints on the coefficient matrix H are necessary. As the coefficient matrix contains the assignment of the participants to the respective patterns, it needs to be ensured that each participant is only assigned to a single pattern, i.e., that each row represents a one-hot-encoded vector. We realized the constraint by adjusting the H-matrix accordingly after each iteration of the solving algorithm instead of incorporating it into the minimization function, which has the advantage of guaranteeing that the constraint is satisfied.

While constraints on the W-matrices are not necessary, they can be used to enforce properties that are tailored to the specific domain. Each column in the W-matrix represents a complete pattern for all 64 syllogisms, which means that chunks of 9 values belong to a single syllogism. Therefore, we normalized each chunk of a column in each iteration of the NMF algorithm to the Euclidean norm in order to obtain results that are more distinct compared to the wider distributions of k-Means. To ensure that the reconstruction remains unaffected, we adjusted the corresponding column of the H-matrix accordingly.

Note that these constraints are not applied to the final results, but after each iteration of the algorithm instead. This ensures that the final result is optimized with respect to the given constraints, which is a major advantage of methods like NMF.

Determining k Given the strong inter-individual differences and noise that become apparent in syllogistic reasoning, it is challenging to determine an optimal (but low) number of clusters since a higher number of clusters would always allow to capture certain individuals better.

To assess this problem, we used a repeated hold-out validation (for different values for k with 1000 iterations each), i.e., we repeatedly divided the data in random subsets (training-set and test-set). Both sets had the same number of participants. We used four metrics to determine the number of clusters and compare the different clustering methods:

The first metric used is the *Inter-Similarity* and assesses the stability of the found patterns with respect to the specific set of participants. If k is too high, patterns might start to represent outliers. In these cases, it is unlikely that the results are stable, as they are likely to jump between different local minima depending on the dataset at hand. Therefore, clustering is performed on both, the training- and the test set. The resulting patterns of both clustering runs are then compared to each other (pattern vs. pattern) using cosine-similarity:

$$sim(w_1, w_2) = \frac{w_1 \cdot w_2}{|w_1||w_2|} \tag{2}$$

Inter-Similarity corresponds to the mean similarity between the patterns obtained from applying clustering to the training- and test set. Since the order of patterns might differ between both runs, the result is only based on the most optimal ordering of the patterns.

The second metric is the *Intra-Similarity*, which has a similar reasoning behind it: If k is too high, patterns might start to be too similar to each other. Therefore, the cosine similarity is used to compare the patterns obtained from a single run of clustering. *Intra-Similarity* is then defined as the maximum similarity between two patterns obtained from the same clustering run. However, the *Intra-Similarity* is unable to distinguish between the occurrence of multiple distinct patterns that are similar to each other and generally less distinct patterns, that have a high similarity because of a more blurry appearance.

For the above-mentioned reasons, we use our third metric, the *Entropy*, which indicates how distinct the pattern is: the more "blurry" a pattern is, the higher the entropy. Therefore, by definition, k-Medoids has a perfect score, as it uses a real participant pattern which always has a distinct response to each task. The entropy for the response distribution for a specific task is calculated as follows:

$$H = -\sum_i p_i * log_2 p_i \tag{3}$$

We use the mean entropy of all tasks of a pattern as the resulting entropy of a pattern.

Finally, we used the *Test-Accuracy*, which is defined as the mean accuracy achieved when using the k patterns obtained from clustering on the training-set as predictors for the participants in the test set. Thereby, the best pattern is selected for each participant.

[1] https://scipy.org

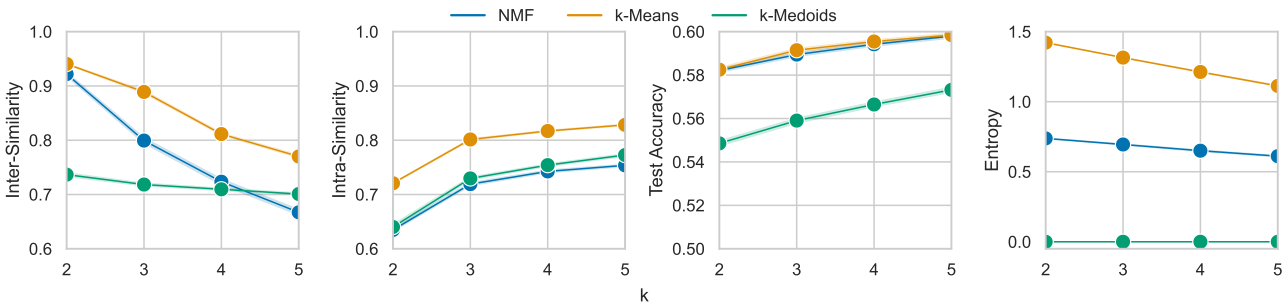

Figure 1: Results of a crossvalidation for kMeans, kMedoids and clustering based on the NMF in terms of inter- and intra-similarity, mean accuracy on the test set and entropy for different numbers of clusters (k).

The results of the metrics for different values of k are shown in Figure 1. For the *Inter-Similarity*, the disadvantage of k-Medoids becomes apparent: The resulting patterns are directly based on the participants, which makes it highly susceptible to changes of the dataset. For NMF and k-Means, a substantial decrease of stability is noticeable with higher levels of k, with k-Means being more robust to the changes overall. However, the downside of k-Means is clearly visible in the *Intra-Similarity*, where its mean-based centroids are substantially less distinct compared to the other methods. For all methods, a substantial change from $k = 2$ to $k = 3$ is apparent, indicating that even a third pattern already leads to a higher similarity between the patterns. However, higher values of k seem to not further increase the similarity to the same extent. This gets confirmed by the *Entropy*, where k-Means also shows to produce less distinct patterns compared to the other methods. This indicates that the worse score of k-Means in *Intra-Similarity* is not due duplicated patterns, but rather an effect of the aggregation. Both, the NMF and k-Means, show an improvement with higher values of k, since the additional clusters allow to build more homogeneous groups. However, as the *Intra-Similarity* indicates, this could also lead to overfitting in the form of almost identical patterns.

For the *Test-Accuracy*, k-Means and NMF show almost the same performance, with k-Medoids falling behind slightly. Also, the differences for varying number of clusters are negligible, suggesting diminishing returns for higher values of k.

Overall, the analysis suggests that a total number of two clusters seems to offer the best trade-off between accuracy and stability. For $k = 2$, the NMF is best suited, since k-Means has a substantially worse *Intra-Similarity* and *Entropy*, while k-Medoids is lacking stability with and therefore also generalizability. Hence, the final patterns (see Figure 2) were obtained with $k = 2$ by using NMF clustering. In the following section, the patterns will be interpreted with respect to their meaning within the domain of syllogistic reasoning.

Interpreting the Patterns

In the following, we will take a closer look on the obtained patterns and the groups of participants that were assigned to

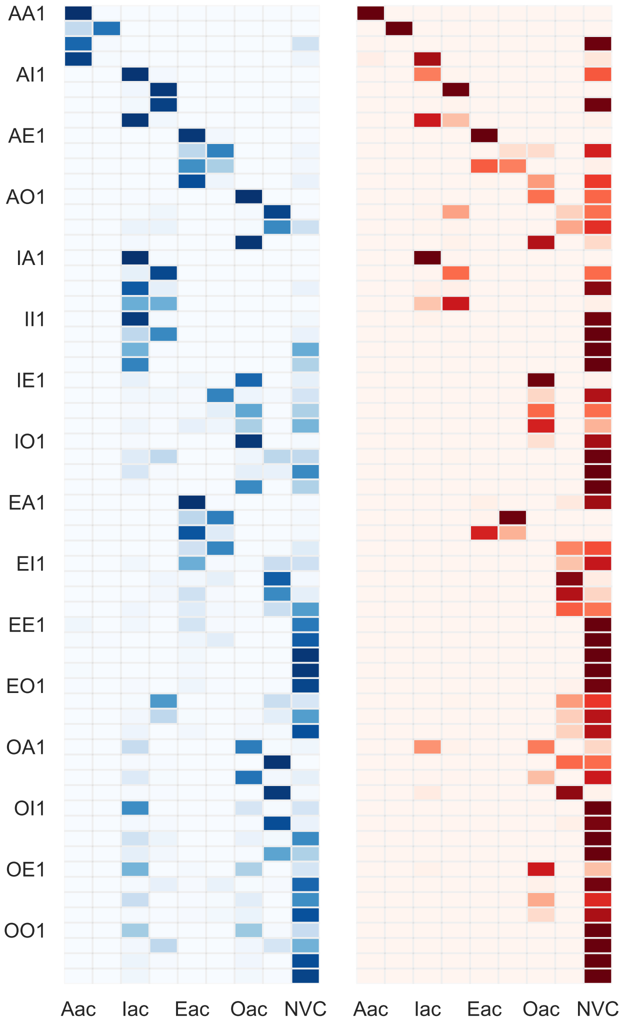

Figure 2: Both response patterns for the 64 syllogistic tasks found by clustering with Non-negative matrix factorization (for $k = 2$). Darker shades denote a higher weight of the respective response.

the respective patterns. For convenience and clarity, we will reference the two groups by *G1* (assigned to the blue pattern) and *G2* (assigned to the red pattern).

When comparing both patterns found by the NMF (see Figure 2), the main difference seems to be revolving around NVC. This is in line with previous analyses, which found the main inter-individual differences to be found with respect to NVC behavior (e.g., Riesterer, Brand, & Ragni, 2020b; Brand et al., 2022). With respect to the logical correctness, G2 also shows a substantially higher correctness ($mean = .68$, $SD = .14$) compared to G1 ($mean = .4$, $SD = .14$), which is expected since NVC is integral for a high correctness due to the high number of invalid syllogisms. While most differences between the patterns are just a shift towards NVC, a slight difference is also apparent for syllogisms with the quantifier *Some not* (O) in the first premise, as responses with the non-negative quantifier *Some* (I) are present for the blue pattern, while - if not NVC - only negative conclusions (*Oac* and *Oca*) are present in the red pattern. Besides these differences, the patterns seem to show identical response patterns.

Given that only two stable patterns were found that differ substantially with respect to their correctness, it is tempting to compare them with dual-processing accounts. Following the idea of dual processes and the respective implementation in mReasoner (Khemlani & Johnson-Laird, 2013), we classify the left (blue) pattern as being more likely to represent a strategy relying on *System 1* (S1), while the right (red) pattern be more frequently engaged in the search for counterexamples and thereby relying on *System 2* (S2). However, it is important to note that the correctness and number of patterns on their own do not corroborate a dual-processing account: Instead, since two stable patterns seem to emerge from the data that mostly differ with respect to NVC, it shows why models for syllogistic reasoning tend to converge to describe inter-individual effects with respect to NVC (i.e., confidences in the Probability Heuristics Model (PHM; Copeland, 2006; Riesterer et al., 2020a), NVC aversion in the model TransSet (Brand, Riesterer, & Ragni, 2020), and the search for counterexamples in MMT (Khemlani & Johnson-Laird, 2013)). Still, assuming a dual-processing account allows us to derive predictions about the groups of participants assigned to the respective patterns by the clustering method: First, it is expected that G2 has a higher response time compared to G1, since relying on the deliberate inferences of S2 should be substantially slower than applying fast-and-frugal heuristics. Second, participants in G2 should show a higher correctness in the Cognitive Reflection Test (CRT), since the test is designed to mislead participants relying on intuition. Furthermore, Need for Cognition (NFC), is also expected to be higher in G2, since participants with high NFC are more likely to engage in tasks that require cognitive effort and deliberative thinking.

With respect to our predictions, we investigated the differences in *Need for Cognition* (NFC) and the correctness in a *Cognitive Reflection Task* (CRT) as well as the mean response time needed for the 64 tasks. The results of the comparison are shown in Table 1. While NFC did not show any significant difference, the CRT differed significantly. With a mean correctness of .47, participants in G1 were substantially more susceptible for the traps of the CRT compared to G2 with a mean correctness of .7. This strengthens the assumption of the dual-process accounts, indicating that G1 relies on a more intuitive process. However, the differences in response times showed that G2 was significantly faster than G1. This contradicts the assumption of a slower, more logical approach using S2, but is in line with previous findings showing faster response times for NVC responses (Brand et al., 2022).

Table 1: Overview and results of a Mann-Whitney-U test between the two groups as assigned by the NMF with respect to Need for Cognition (NFC), Cognitive Reflection Task correctness (CRT) and the mean response times (RT). Factors showing significant differences (with Bonferroni corrected $\alpha = 0.0167$) are written in bold.

	Mean		SD		U	p
	G1	G2	G1	G2		
NFC	4.65	4.73	.9	.84	1224.5	.536
CRT	.47	.7	.29	.28	747.0	$< .001$
RT	15803	13468	5969	6610	1697.0	.001

Finally, we checked how well the participants would be classified based on the NFC and CRT. To this end, we reassigned the participants to the two patterns based on their NFC and CRT scores (using the median as a threshold). Subsequently, the accuracy of the respective pattern in predicting the participant's responses was calculated for each participant. Additionally, we included the original assignment as obtained from the NMF (*Fit*) and a post-hoc optimal assignment maximizing accuracy. Furthermore, the accuracy of the Most-Frequent Answer (MFA) strategy was added as a baseline. The MFA could thereby be understood as the result of a clustering with $k = 1$, making it useful to assess the additional gain by having an additional pattern. The results are depicted in Figure 3.

As expected from the previous analysis, NFC could not be used as an assignment strategy, even decreasing the accuracy (0.514) below the level of the MFA (0.552). However, the CRT only managed to improve the accuracy slightly (0.555), illustrating that a significant factor does not necessarily translate into being a powerful predictor on the level of individual response predictions. Finally, both data-driven assignments achieve an almost identical performance (0.599) which is a substantial improvement over both, the CRT and the MFA.

Discussion

The key goal of this article was to find and investigate stable patterns of human syllogistic reasoning behavior, which could be considered iconic for the task. Our analysis shows that first, only two patterns can be identified robustly, and sec-

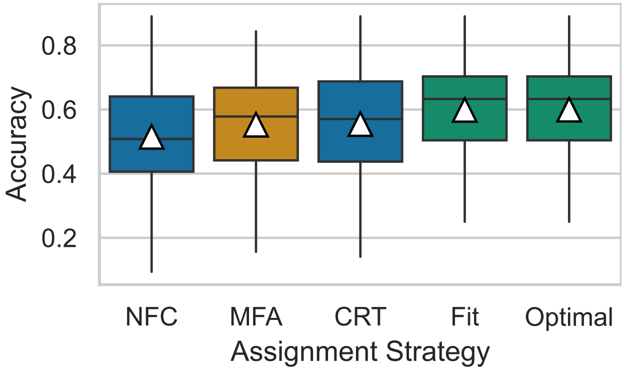

Figure 3: Accuracy achieved when comparing the individual patterns to the two iconic patterns for different assignment strategies. Individual traits are shown in blue, data-driven assignments in green. As a baseline, the most-frequent answer (MFA; orange) is added. Triangles denote the mean accuracy.

ond that these patterns differ most predominantly in terms of the frequency of relying on NVC as a conclusion option.

The composition of the two identified patterns explains why most of the proposed models in syllogistic reasoning research (e.g., see Khemlani & Johnson-Laird, 2012) converge to a similar distinction between NVC-friendly and an NVC-averse participants in their response predictions. Instead, their key differences are mostly in the explanation of why these patterns emerge in syllogistic reasoning.

The fact that two patterns are found specifically seems to corroborate the dual-processing account assumptions underlying the search for counterexamples in MMT. This is further reinforced by the fact, that the reasoners associated with the more correct pattern also score high on the CRT, which is designed to assess the affinity of reasoning in a deliberative and logically correct manner. However, the observed response times associated with the patterns are contradictory to what is posited by the theory: the logically correct pattern is associated with faster instead of slower reaction times. Additionally, the CRT is known to correlate with various measures of cognitive ability (Frederick, 2005), which could also explain the a higher performance on syllogistic tasks. As a side-note, the marginal improvement achieved by using the CRT as an assignment strategy illustrated a pitfall in cognitive modeling: Even highly significant factors due not necessarily translate well to the level of predictors for individual patterns.

The results shown in this article raise the question if traditional modeling of syllogistic reasoning behavior has hit a dead end or will hit it soon. As models converge to the same patterns and only differ in their sets of explanatory assumptions, new experiments need to be designed and datasets acquired to more specifically investigate the validity or falsity of the underlying assumptions. One step towards this goal could be to integrate more auxiliary information about individuals for example via extended psychological test batteries. This could also make it possible to find additional patterns more nuanced to smaller sub-populations of participants. Furthermore, the explanatory component of models and their underlying theories will be of greater importance, since a model comparison purely based on the general patterns will not suffice for a meaningful distinction between the models' capabilities. Instead, deriving specific hypotheses tailored to test certain assumptions of the model will become necessary.

On a technical level, our work showed that clustering, especially with flexible approaches like Non-Negative Matrix Factorization, can help to uncover expressive iconic patterns in human reasoning data. Paired with the proposed metrics, which allow to assess the robustness of the found patterns in domains where large inter-individual differences are to be expected, these approaches are valuable assets in cognitive modellers' toolkits, irrespective of the domain of interest.

Acknowledgements

This research was supported by the German Research Foundation, DFG (Grant RA1934/5-1 and RA1934/8-1) within the SPP 1921 "Intentional Forgetting" and by the Saxony State Ministry of Science and Art (SMWK3-7304/35/3-2021/4819) excellence initiative "Productive Teaming" on the basis of the budget passed by the deputies of the Saxony state parliament.

References

Aggarwal, C. C. (2015). Cluster analysis. In *Data mining: The textbook* (pp. 153–204). Cham: Springer International Publishing.

Bara, B. G., Bucciarelli, M., & Johnson-Laird, P. N. (1995). Development of syllogistic reasoning. *The American Journal of Psychology*, *108*(2), 157.

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.

Brand, D., Riesterer, N., & Ragni, M. (2022). Model-based explanation of feedback effects in syllogistic reasoning. *Topics in Cognitive Science*, *14*(4), 828-844.

Cacioppo, J. T., & Petty, R. E. (1982). The need for cognition. *Journal of personality and social psychology*, *42*(1), 116.

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.

Dames, H., Klauer, K. C., & Ragni, M. (2022). The stability of syllogistic reasoning performance over time. *Thinking & Reasoning*, *28*(4), 529-568.

Evans, J. S. B. T. (2008). Dual-processing accounts of reasoning, judgment, and social cognition. *Annual Review of Psychology*, *59*(1), 255–278.

Frederick, S. (2005). Cognitive reflection and decision making. *Journal of Economic perspectives*, *19*(4), 25–42.

Gilhooly, K. J., Logie, R. H., Wetherick, N. E., & Wynn, V. (1993, jan). Working memory and strategies in syllogistic-reasoning tasks. *Memory & Cognition*, *21*(1), 115–124.

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.

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.

Kim, H., & Park, H. (2008). Nonnegative matrix factorization based on alternating nonnegativity constrained least squares and active set method. *SIAM Journal on Matrix Analysis and Applications*, *30*(2), 713–730.

Kim, J., & Park, H. (2008). *Sparse nonnegative matrix factorization for clustering* (Tech. Rep.). Georgia Tech.

Lawson, C. L., & Hanson, R. J. (1995). *Solving least squares problems*. Philadelphia: SIAM.

Morgan, J. J., & Morton, J. T. (1944). The distortion of syllogistic reasoning produced by personal convictions. *The Journal of Social Psychology*, *20*(1), 39–59.

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). Feedback influences syllogistic strategy: An analysis based on joint nonnegative matrix factorization. In *Proceedings of the 18th international conference on cognitive modeling* (pp. 223–228).

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

Rips, L. J. (1994). *The psychology of proof: Deductive reasoning in human thinking*. Mit Press.

Störring, G. (1908). *Experimentelle untersuchungen über einfache schlussprozesse*. W. Engelmann.

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.