On the Matter of Aggregate Models for Syllogistic Reasoning: A Transitive Set-Based Account for Predicting the Population

Daniel Brand* (daniel.brand@cognition.uni-freiburg.de)
Nicolas Riesterer* (riestern@cs.uni-freiburg.de)
Marco Ragni (ragni@cs.uni-freiburg.de)
Cognitive Computation Lab, Georges-Khler-Allee 79
79110 Freiburg, Germany

Abstract

Recent work in modeling human syllogistic reasoning claimed that heuristic approaches perform worse in accounting for experimental data than more comprehensive representations of cognition. We show that this observation might have been due to a misconception of the goals heuristics are often developed for: representing a specific psychological phenomenon or reflecting individual inference strategies. To demonstrate the performance of heuristic models, we introduce a novel model for syllogistic reasoning fundamentally based on transitivity. By evaluating it based on predicting the most frequent answer, i.e., the response most often selected by participants, we show that this model is able to outperform the current state of the art, demonstrate the promising role of transitive inferences in syllogistic reasoning, and discuss its implications for modeling individual reasoners instead of populations.

Keywords: syllogistic reasoning; predictive modeling; heuristics; transitivity

Introduction

Syllogistic reasoning is, next to conditional and relation reasoning, one of the core domains of human reasoning research (Evans, 2002). Syllogisms are quantified statements of the form "All pilots are painters, Some painters are divers" consisting of two premises which are constructed by relating two terms A-B (i.e., pilots-painters), and B-C (i.e., painters-divers) via one quantifier out of "All, Some, No, Some ... not" (for additional background information see Khemlani & Johnson-Laird, 2012). Depending on the order of terms in the premises, syllogisms can be classified into four figures:

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

The goal of syllogistic reasoning tasks is to use the information of the premises which are related to each other via the middle term B in order to draw a conclusion about the end terms A, C by using one of the quantifiers mentioned above or infer "No Valid Conclusion" (NVC) if there is none. In total, by considering all combinations of quantifiers and figures, there are 64 distinct syllogistic problems with nine possible conclusions causing the domain to be well-defined and accessible for cognitive modeling endeavors. To increase readability of the syllogistic problems, quantifiers will be represented in accordance to their traditional latin abbreviations (originating from "affirmo" and "nego") by an uppercase letter for the remainder of the article:

All	Some	No	Some ... not
A	I	E	O

Syllogistic problems are encoded by specifying these quantifier encodings as well as the figural identifier (e.g., AI1 for "All A are B; Some B are C").

Research shows that human syllogistic inferences differ substantially from classical logics (Wetherick & Gilhooly, 1995). Over the course of the last decades, multiple statistical effects and psychological phenomena were identified and used to formulate hypotheses and theories about mental representations and inferential mechanisms used when reasoning over syllogisms. Traditionally, analyses of syllogistic models are based on aggregated data resulting in models being evaluated in terms of their capability to capture an "average" reasoner. As an example, the authors of a meta-analysis (Khemlani & Johnson-Laird, 2012) relied on hits, correct rejections, and correct predictions to quantify the match between model predictions and experimental data. Their results showed that no satisfactory ordering of model performances could be identified as all theories exhibited distinct strengths and weaknesses with respect to the evaluation metrics.

In this paper, we introduce a novel model for syllogistic reasoning — TransSet — which is based on a heuristic use of transitive inferences. We evaluate the model by focusing on the ability to predict the most frequently given answer (MFA) to a syllogism. This reflects the response given by the "average" reasoner, which lies at the center of population-based analyses. The model's performance is discussed and compared to the state of the art models in cognitive modeling of syllogistic reasoning. Additionally, since group-level results do not necessarily generalize to the individual level (Molenaar, 2004; Fisher, Medaglia, & Jeronimus, 2018), we investigate the transferability of the results to the level of individuals.

The structure of the remainder of the article is as follows. First, we introduce related literature on cognitive modeling in the field of syllogistic reasoning as well as on statistical effects and psychological phenomena we base our model

*Both authors contributed equally to this manuscript.

on. Second, we give details about the model's computational principles along with an overview of the responses it is able to predict. Third, we perform the predictive analysis of the state of the art and our newly proposed model. Finally, the implications of the results are discussed and directions for future work are suggested.

Related Work

Developing accurate models to explain and predict human responses which differ greatly from classical logics (Wetherick & Gilhooly, 1995) has been a core focus of syllogistic reasoning research in the past decades. Currently, there exist at least twelve cognitive theories attempting to give explanations about the inferential mechanisms inherent to human cognition by relying on a multitude of different methodological foundations (Khemlani & Johnson-Laird, 2012). The authors of a recent meta-analysis (Khemlani & Johnson-Laird, 2012) proposed a classification of the existing theories into *heuristic theories* largely based on simple explanations for differences from classical logics, *formal rule theories* mainly proposing logic-based inference mechanisms, and *theories based on diagrams, sets, or models* which focus on mental representation of information and corresponding inferential operations.

Recently, an increasing effort has been made to turn abstract and often underspecified cognitive theories of syllogistic reasoning into computational models allowing for an assessment of predictions. In their meta-analysis, Khemlani and Johnson-Laird (2012) compiled prediction tables for most of the cognitive theories resulting in an analysis showing that the existing theories feature distinct predictive properties with respect to hits, correct rejections, and correct predictions. In consequence, no clear ranking of the models' predictive qualities could be determined.

One minor result of the meta-analysis was that heuristic models generally perform worse than more elaborate comprehensive accounts which try to give more detailed explanations about cognition by tying into mental representation, memory, or other components of the human mind (for an example, see the mental models theory, MMT, Johnson-Laird, 1983). However, as recent work shifting the focus of analysis to predicting responses could show, the poor performance of heuristics might have been due to a mismatch of modeling purpose and intent. Since heuristics do not aim at explaining the general population but attempt to formalize specific strategies which may be applied by certain individuals, caution needs to be exercised when analyzing comparative performance evaluations. Indeed, recent work combining heuristics to form a composite portfolio model demonstrated a substantial improvement in performance when leveraging strengths while avoiding weaknesses of specific heuristic accounts (Riesterer, Brand, & Ragni, 2018). A conclusion of this work is that heuristic models should not be underrated in general cognitive modeling. While potentially unsuitable as comprehensive accounts of human cognition, they might be able to reflect strategies and mechanisms employed by individuals. Because of this they can serve as promising test benches to investigate the role of the numerous statistical effects and psychological phenomena uncovered.

A fundamental concept of human reasoning that has been extensively investigated is transitivity. In the domain of reasoning in particular, the term *pseudo-transitive fallacy* was introduced to describe the phenomenon that human reasoners are prone to drawing transitive inferences even if logically unwarranted (Goodwin & Johnson-Laird, 2008). Some reasoners also assumed transitivity and symmetry when presented with a completely unknown relation (Tsal, 1977).

In the following, we rely on transitivity to develop a novel heuristic model of human syllogistic reasoning which is based on transitive chains of information. The idea to explain syllogistic reasoning based on transitive effects is not new. Guyote and Sternberg (1981) introduced a model which represents information as pairs and integrates set relations via rules applied to transitive chains of information. The difference to what we propose is that transitivity is used as the driving factor for reasoning. Our model assumes transitivity to serve a heuristic purpose allowing humans to avoid relying on higher-level reasoning processes.

A Transitive Model

A major part of the inferences that are drawn on a regular basis in daily life are transitive (e.g., A is bigger than B, B is bigger than C, therefore A is bigger than C). Usually, these kinds of inferences are easy for human reasoners to draw. On the other hand, tasks that look like transitive inference tasks at first glance, when in reality they are not, are prone to errors originating from an unwarranted use of transitivity. It can be assumed that the simplicity and familiarity of transitive tasks plays a major role for this kind of fallacy.

In the following we propose a heuristic model for syllogistic reasoning based on the principle of transitivity. The main assumption of the heuristic model is that some human reasoners try to circumvent a fully fleshed-out inference process by trying to apply simple rules for patterns they are familiar with from transitive inferences. Here lies the major difference to the transitive-chain theory (Guyote & Sternberg, 1981), which is a theory of the human reasoning process instead of a heuristic which might be used by some reasoners to avoid in-depth inference processes by applying shallow transformations to obtain familiar patterns.

The general process of TransSet is sketched in Figure 1. Its first step focuses on determining the direction of the syllogism by looking for a transitive pattern A-B-C or C-B-A. Such patterns can be found directly for syllogisms with figure 1 and 2, corresponding to a path from A to C and from C to A, respectively.

For Figure 3 and 4 this process fails, which leads to an NVC response in most cases. In some cases, however, a path can be constructed by changing the direction of one of the premises: Figure 3 syllogisms consist of two premises featur-

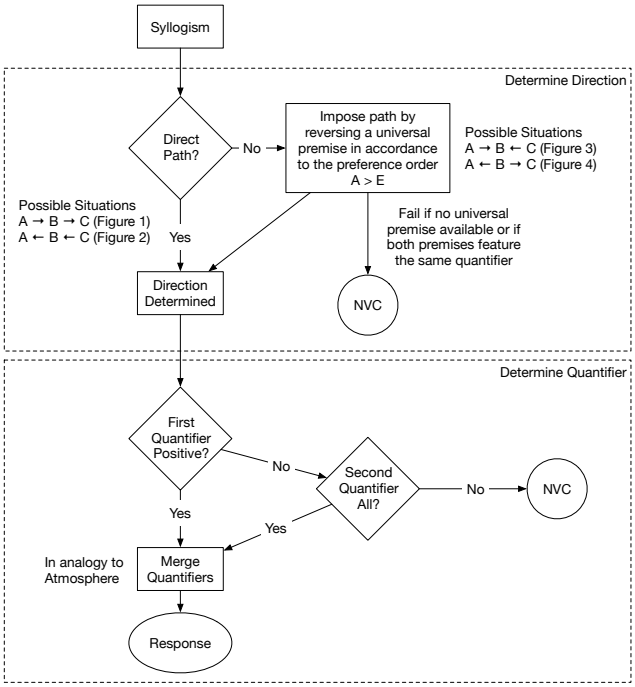

Figure 1: Flow of the TransSet model. First, the direction is determined by extracting a transitive path from the premise information. Second, the quantifier is determined by merging quantifiers. In case of failures resulting from insufficient information or disconnected premises, NVC is generated.

ing paths to B, while Figure 4 syllogisms yield paths starting from B. In both cases it has to be decided which categorical set (A or C) should be put in the place of B. At first glance, it might seem reasonable to choose the set of elements that has the most informative quantifier as it is able to "compensate" for the uncertainty introduced by changing the premise direction. We consider the universal quantifiers A and E as "informative", since they make statements about all elements in a set. The TransSet model therefore changes the direction of the premise with the most informative universal quantifier, if there is a single "most informative" universal quantifier (with an ordering of $A > E$). In case of ties or a lack of informative quantifiers, the process fails and returns NVC. Note, that the change of direction requires the assumption of symmetry, which is logically invalid for the quantifier A. The occurrence of this deviation from classical logic in human reasoning behavior is also a core concept of the conversion theory (Revlis, 1975).

As soon as a path is obtained, the task can be solved by propagating the starting set of elements along the path (while applying the quantifiers). For example, considering syllogism AI1 (All A are B, Some B are C), a set consisting of all A is propagated to B, where it is filtered by the second quantifier on its path from B to C, reducing the set to "Some A". Therefore, the conclusion would be "Some A are C", which is logically invalid. It is important to note that the process of set propagation yields the same conclusion quantifiers as the atmosphere theory (Wetherick & Gilhooly, 1995), but also predicts the direction of the answer: a path from A to C naturally corresponds to an answer with the direction $A \rightarrow C$. The resulting predictions are in line with the figural effect (Johnson-Laird, 1983).

The propagation, however, does not succeed in all cases. When the set obtained after filtering by the first quantifier is empty, traversing the transitive path is no longer possible. For example, when considering syllogism EI1, the set after the path $A \rightarrow B$ would be empty, as there are no elements from A that are also B. It is therefore not possible to integrate the second quantifier, as the set cannot be reduced any further. This leads to the NVC response, since the endpoint of the path cannot be reached. An exception to this can occur if the second quantifier is A: because A does not require any filtering, it corresponds to simply passing the set ahead, which prevents the path from breaking. Note, that this failure of the propagation induces an asymmetry regarding the quantifier which is not generally assumed in heuristic models: since it can only happen if the first processed quantifier leads to an empty set, syllogism EI1 and IE2 are affected but IE1 and EI2 are not.

The TransSet model is a heuristic model. As such, it only describes a single heuristic strategy assumed to be used by some human reasoners for syllogistic reasoning. Therefore, we used the heuristic in a strictly deterministic setup, where a single prediction for each syllogism was generated according to the procedure described above. The resulting predictions are shown in Table 1.

Analysis

The following analysis is based on the dataset and models reported by Khemlani and Johnson-Laird (2012). Additionally, we included a separate analysis on a dataset of 139 reasoners obtained from a web experiment conducted on Amazon Mechanical Turk which was published as part of the benchmarking framework CCOBRA[1]. This second dataset is not only included to extend the size of the evaluation dataset, but also because it contains unaggregated responses to syllogistic problems which can be used to assess a model's capability to account for individual reasoners. All files related to the following analyses are available on GitHub[2].

MFA Assessment

First, we investigate how accurately models are able to predict the MFA by comparing the set of possible predictions for a given syllogism with the most frequently selected response in the data.

Figure 2 depicts the results of this evaluation based on two different metrics. The left plot presents the proportion of syllogistic problems which feature an MFA response that is contained in the set of possible predictions by the respective model. The obtained values differ substantially be-

[1] https://github.com/CognitiveComputationLab/ccobra
[2] https://github.com/Shadownox/iccm-transset

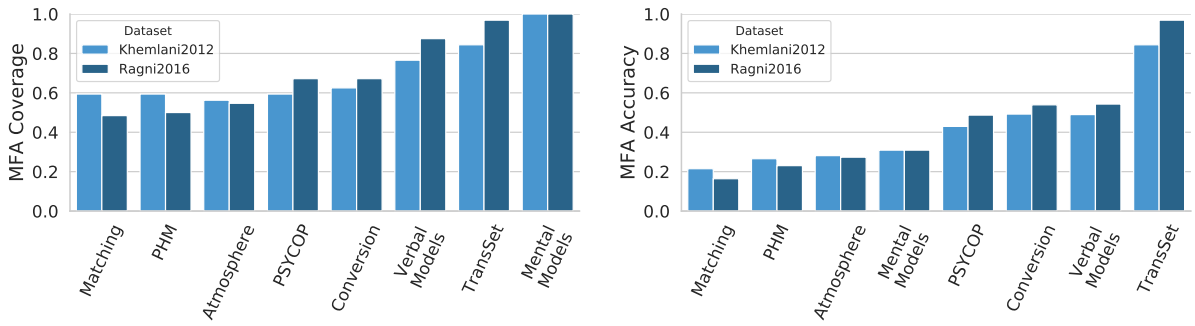

Figure 2: Analysis of predictive performances based on MFA comparison. Left plot depicts proportion of syllogistic problems where at least one of the possible model predictions matches the MFA. Right plot depicts accuracy of predicting MFA (discounted for multiple possible conclusions).

Table 1: Predictions of the TransSet model.

Syllogism	Prediction	Syllogism	Prediction
AA1	Aac	EA1	Eac
AA2	Aca	EA2	Eca
AA3	NVC	EA3	Eac
AA4	NVC	EA4	Eca
AI1	Iac	EI1	NVC
AI2	Ica	EI2	Oca
AI3	Ica	EI3	Oca
AI4	Iac	EI4	NVC
AE1	Eac	EE1	NVC
AE2	Eca	EE2	NVC
AE3	Eca	EE3	NVC
AE4	Eac	EE4	NVC
AO1	Oac	EO1	NVC
AO2	Oca	EO2	NVC
AO3	Oca	EO3	NVC
AO4	Oac	EO4	NVC
IA1	Iac	OA1	Oac
IA2	Ica	OA2	Oca
IA3	Iac	OA3	Oac
IA4	Ica	OA4	Oca
II1	Iac	OI1	NVC
II2	Ica	OI2	Oca
II3	NVC	OI3	NVC
II4	NVC	OI4	NVC
IE1	Oac	OE1	NVC
IE2	NVC	OE2	NVC
IE3	Oac	OE3	NVC
IE4	NVC	OE4	NVC
IO1	Oac	OO1	NVC
IO2	NVC	OO2	NVC
IO3	NVC	OO3	NVC
IO4	NVC	OO4	NVC

tween models. While heuristics such as Matching, the Probabilistic Heuristic Model (PHM), or Atmosphere only contain the MFA response in less than 60% of syllogistic problems, model-based approaches such as the Mental Models Theory (MMT) or Verbal Models are able to achieve above 80%. These observations are in line with the results obtained by Khemlani and Johnson-Laird (2012). However, despite its fundamentally heuristic principles, TransSet is capable to compete with the most performant state of the art models arriving at MFA coverage proportions of above 80% demonstrating that heuristic principles are not generally inferior to more comprehensive models.

A shortcoming of this type of coverage-based analysis is that it ignores the size of the sets of possible model predictions. However, since the more responses a model is allowed to include the higher the possibility is to cover the MFA, models need to be penalized for unnecessary predictions. This is presented in the right plot of Figure 2 which assigns a score of $1/|P_s|$ if the MFA is contained in the prediction set P_s thereby introducing a penalty factor linear in the number of possible predictions. As a result, a model is given a score of 1 if it does not include other responses apart from the MFA for all syllogisms and lower scores if unnecessary conclusions are predicted. For example, the mental models theory captures the MFA "Aac" for syllogism "AA1" in its prediction set $\{\text{Aac}, \text{Aca}, \text{Ica}\}$. As a result it is assigned a score of $1/3$.

This plot draws a different picture of model performances. It shows that when discounting scores based on the number of predictions, performances drop considerably. MMT and Verbal Models which dominated the coverage analysis (left plot) drop substantially due to the fact that they include up to five of the nine possible conclusions in their prediction sets. TransSet on the other hand remains unchanged since it only allows a single prediction to each syllogistic problem.

Put together, both plots demonstrate that the high levels of accuracy achieved by some models (Mental Models, Verbal Models) are mainly due to their large numbers of predicted responses. When compared to TransSet, however, it becomes apparent that complex and potentially parameterized mod-

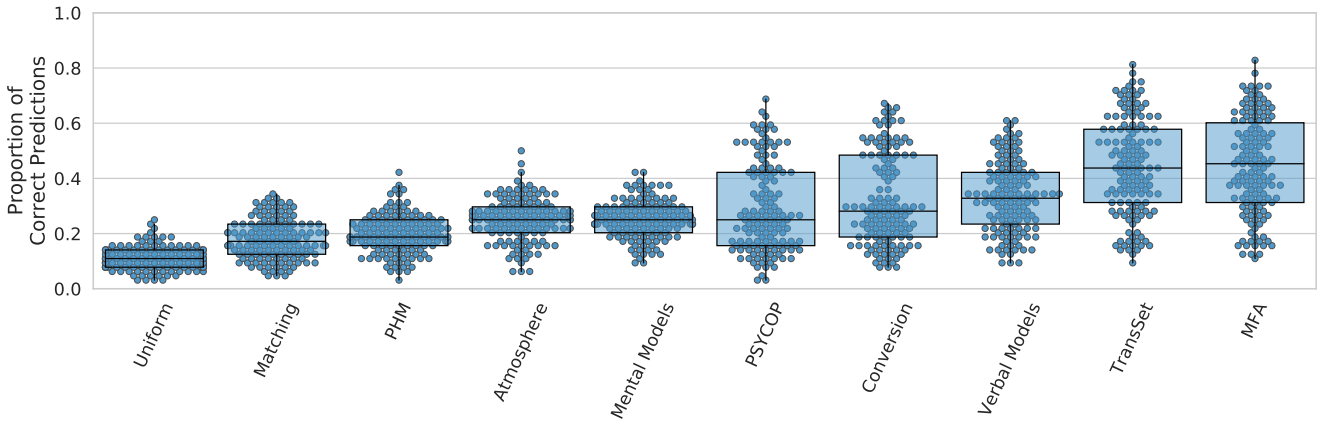

Figure 3: Evaluation of the predictive accuracies of syllogistic models. Boxplots denote medians, inter-quartile range (IQR) as well as whiskers extending to the last data point within a distance of 1.5 times the IQR from the edge of the boxes. Points indicate the accuracy for a specific individual reasoner. Accuracies on the CCOBRA 139 participant dataset of the state-of-the-art models computed from the predictions as reported by Khemlani and Johnson-Laird (2012) are depicted alongside TransSet and two baseline models: "Uniform" corresponds to using a uniform distribution to randomly select an answer and "MFA" reflects the most-frequent answer strategy.

els are unnecessary for predicting aggregated responses. The heuristic principles based on notions of transitivity resulting in a single response suffice to achieve state-of-the-art performance. Note that this analysis focusses on the ability of predicting responses, whereas some models might also be able to provide additional estimates (e.g., reaction times), which are out of scope for the present article.

Individual Match

The results show that TransSet is able to cover a majority of the most frequently given answers and is able to account for populations of reasoners. The pressing question is how relevant MFA is for capturing the variety of strategies that are employed by human reasoners. Put differently, it would be insightful to see whether reasoners differ with respect to their strategies and whether MFA, and accordingly TransSet as its approximation, is a suitable representation for the majority of reasoners.

The second analysis reported in the following therefore shifts the focus towards an assessment of the capability of models to account for the strategies employed by individual human reasoners. In particular, we investigate the match between model predictions and individual responses of the 139 participants contained in the CCOBRA dataset.

Figure 3 depicts the evaluation output obtained from the benchmarking framework CCOBRA. The image depicts the accuracy of individual models when predicting responses for individual reasoners (dot swarm). The box plots present an aggregated representation of these exact results. The image shows that models achieve low predictive accuracies across the board with TransSet surpassing the current state of the art. The swarm plots show that variances of accuracies differ greatly between models. While models on the lower end of the spectrum produce accuracies between 0% and 40%, TransSet is able to predict up to 80% of an individual's responses correctly.

There are two sides to the results depicted here. On the one hand, it is interesting to see that some of the models are able to successfully predict most of the responses for at least a small part of the population. On the other hand, it shows that not even MFA is able to adequately cover the majority of people. This demonstrates that syllogistic model evaluation solely on aggregated data is severely limited and not necessarily generalizable to individuals. This puts the general goals of cognitive modeling into perspective. A model that claims to reflect cognitive processes or general phenomena of nature in a suitable manner should always be able to achieve high levels of predictive accuracy. If we assume reasoners to rely on a large number of independent strategies this would correspond to models being able to match certain individuals well while completely failing to capture others. This is often the case for heuristics, since the phenomena or cognitive fallacies they are constructed on are only applicable to a subset of individuals in the population. Models accounting for general principles, on the other hand, should generally show a smaller variance in coverage of individuals since the principles should be prevalent in all responses to some degree.

Discussion

In this article we introduced TransSet, a novel model for predicting human syllogistic reasoning. Drawing from the statistical effects and psychological phenomena of the recent literature, TransSet is capable of competing with state-of-the-art models by relying on deterministic and heuristic principles only. When discounted for the number of possible predictions a model generates for a syllogism, TransSet is able to

achieve a coverage of MFA of above 80% resulting in an improvement of about 20% over the state of the art as reported by Khemlani and Johnson-Laird (2012).

The main conclusions of this article are twofold. First, we demonstrate that complex parameterized models are not required when aiming for predicting an "average" reasoner, i.e., aggregated data. TransSet, which generates a single deterministic response to each syllogism is not only competing with but outperforms the state of the art when discounting for the number of possible responses. Second, the evaluation of predictive accuracy on individuals highlights that no existing model is able to adequately reflect the reasoning strategy employed by the majority of participants. In order to not only account for a select few reasoners but for a wide variety of individuals, adaptive models tuned to the inferential mechanisms of specific reasoners are required. This, however, remains an open challenge for future work.

TransSet's performance is made possible because it incorporates effects and phenomena uncovered in empirical research. As such it is comprised of ideas found in other models (e.g., transitivity and illicit conversion) and as such can be understood as a superset of models. The fact that a simple model based on heuristic principles is able to outperform the state of the art illustrates the potential that remains in the field. Especially when moving beyond models for aggregated data, the adaptability of parameterized models to individual inferential mechanisms will allow for an even better understanding of cognition and consequently for the development of more accurate models.

Human syllogistic reasoning is far from being solved. In addition to outperforming the state of the art in the aggregate case, TransSet demonstrates a performance that suggests that its underlying concepts form a plausible reasoning strategy for at least some individuals. The heuristic use of transitivity has therefore proven to be a powerful mechanism for explaining human syllogistic reasoning performance and might suggest connections to related results from cognitive science indicating that humans are generally likely to draw transitive conclusions even when they are unjustified (Goodwin & Johnson-Laird, 2008). It remains to be seen if the model can be transferred to other domains featuring transitive properties successfully (e.g., spatial-relation or conditional reasoning). Currently, we only focus on a direct extraction of general output predictions from the model. Future work will focus on two directions: First, we will investigate possible parameterizations allowing the model to fine-tune itself to individual human reasoners. Second, we will investigate further properties of the reasoning process such as reaction times or its connection to the psychological phenomena of syllogistic reasoning.

Acknowledgements

This paper was supported by DFG grants RA 1934/3-1, RA 1934/2-1 and RA 1934/4-1 to MR.

References

Evans, J. S. B. T. (2002). Logic and human reasoning: An assessment of the deduction paradigm. *Psychological Bulletin*, *128*(6), 978–996.

Fisher, A. J., Medaglia, J. D., & Jeronimus, B. F. (2018). Lack of group-to-individual generalizability is a threat to human subjects research. *Proceedings of the National Academy of Sciences*, *115*(27), E6106–E6115.

Goodwin, G. P., & Johnson-Laird, P. (2008). Transitive and pseudo-transitive inferences. *Cognition*, *108*(2), 320–352.

Guyote, M. J., & Sternberg, R. J. (1981). A transitive-chain theory of syllogistic reasoning. *Cognitive Psychology*, *13*(4), 461–525.

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

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

Molenaar, P. C. M. (2004). A manifesto on psychology as idiographic science: Bringing the person back into scientific psychology, this time forever. *Measurement: Interdisciplinary Research & Perspective*, *2*(4), 201–218.

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

Riesterer, N., Brand, D., & Ragni, M. (2018). The predictive power of heuristic portfolios in human syllogistic reasoning. In F. Trollmann & A.-Y. Turhan (Eds.), *Proceedings of the 41st German Conference on AI* (pp. 415–421). Berlin, Germany: Springer.

Tsal, Y. (1977). Symmetry and transitivity assumptions about a nonspecified logical relation. *Quarterly Journal of Experimental Psychology*, *29*(4), 677–684.

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