Bachelor Thesis

Evaluation of Model Parameter Configuration Based on Feedback Data

Cognitive Computation - Albert-Ludwigs-University Freiburg

Nils Engelhardt
Scheffelstraße 32
79102, Freiburg im Breisgau
Matriculation Number: 3741160

Advisors:	M. Sc. Nicolas Riesterer, M. Sc. Daniel Brand
First Examiner:	Prof. Dr. Dr. Marco Ragni
Study Course:	Computer Science
Period:	Winter Semester 2020/2021
Date:	February 1, 2021

Zusammenfassung

Menschliches Denken durch rechnerische Ansätze simulieren zu können, hängt in erster Linie von unserem Verständnis ab, wie Urteile und Entscheidungen getroffen werden. In dieser Studie werden drei verschiedene parametrisierte Rechenmodelle analysiert, die das menschliche Denken bei der Lösung von syllogistischen Problemen abbilden sollen. Ziel ist es, die besten Parameter Konfigurationen der Modelle zu identifizieren, um bevorzugte Strategien des menschlichen Denkens abzuleiten, wenn sie mit Syllogismen konfrontiert werden. Zu diesem Zweck wurde das menschliche Entscheidungsverhalten durch Feedback beeinflusst. Aus der Analyse dieser Parameter können die bevorzugten Methoden der menschlichen Entscheidungsfindung abgeleitet werden. Einige Modelle legen nahe, dass Menschen bei ihren Entscheidungen skeptischer werden, wenn sie Feedback darüber bekommen, ob ihre Entscheidung richtig oder falsch war. Im Gegensatz dazu zeigt das Modell TransSet, das am besten den durchschnittlichen Menschen zu beschreiben scheint, eine geringere Aversion der Menschen bezüglich der Antwort einer „ungültigen Schlussfolgerungen“ (NVC), wenn Feedback gegeben wird, was als zunehmendes Vertrauen in diese *ungewohnte* Antwort interpretiert werden kann.

Abstract

Simulating human reasoning by computational approaches is first of all dependent on our understanding of how judgments and choices are made. In this study three different parameterized computational models which aim to represent human reasoning when solving syllogistic problems are analyzed. The goal is to identify the parameter configurations which describe the average human reasoner best in order to infer preferred methods of human thinking when challenged with syllogisms. For this reason the human reasoning behaviour was altered by giving feedback. Each model indicates that the average person has a tendency to think intuitively and is prone to making logically incorrect decisions. Moreover, some models suggest that people become more sceptical in their decisions upon feedback when it comes to solving syllogistic problems. In contrast, the best performing model TransSet indicates less reluctance by the participants in answering ”no valid conclusion” (NVC) when feedback is given which can be interpreted as increasing confidence towards this unfamiliar response.

Contents

List of Figures

List of Tables

1 Introduction

Although humans are not irrational, they often make false decisions or inaccurate judgements. In the 1970s for instance, Tversky and Kahneman describe in their seminal work how people employ heuristics in situations of uncertainty (Tversky & Kahneman, 1974). Heuristics are simple processes, which are used to find a likely answer that is not necessarily logically correct. Thus, to better understand human reasoning and to comprehend mental processes, one must not be deluded by what would be logically correct but rather include psychological and cognitive aspects into consideration. Ever since, research on human reasoning has been continued and refined in order to understand its strategies and principles. One way to investigate human reasoning is by challenging people with syllogistic problems.

Syllogistic reasoning has been one of the major fields in human reasoning research for decades. First investigations were made by Störring at the beginning of the 20th century (Störring, 1908). A syllogistic problem traditionally contains two quantified premises (*all*, *some*, *no*, *some...not*), which state a relation between *A and B* and *B and C*. Given those premises, the reasoner is asked to derive a conclusion regarding the relation between the end-terms *A and C* or vice versa. E.g., given the premises 1) *all divers are adventurers* and 2) *all adventurers are treasure hunters*, one may conclude that *all divers are treasure hunters*. However, sometimes a conclusion is impossible to determine. If *all divers are adventurers* and *some adventurers are treasure hunters*, it is impossible to derive a logically correct conclusion, because just a subset of *adventures* are *treasure hunters*. Hence, the relationship between *divers* and *treasure hunters* could be anything. In this case, the reasoner is allowed to give a *No Valid Conclusion* (NVC) response. The order of terms within the premises may vary, what leads to four different classifications (figures) of syllogistic problems (Khemlani & Johnson-Laird, 2012) illustrated in Table 1.1.

Table 1.1: Figures of syllogistic problems.

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

To improve readability, the quantifiers have been given abbreviations: A for *all*, I for *some*, E for *no* and O for *some...not*. Thus, *all divers are adventurers* and *all adventurers are treasure hunters* can be encoded as *AA1*: The quantifier is *all* in both premises and the terms are ordered as illustrated for Figure 1 shown in Table 1.1. Combining all the quantified premises with terms A, B and C, there are 64 different syllogistic tasks in total which can be solved using a set of eight quantified responses and NVC. However, just one response at a time is usually allowed to be given.

According to Woodworth and Sells, human reasoners tend to give logically incorrect responses for syllogistic problems because humans often rely on their quick, intuitive and effortless thinking rather on what would be logical (Woodworth & Sells, 1935). In order to understand and reconstruct human reasoning in this matter, there is research on computational approaches, which potentially are capable of simulating the mental processes occurring while solving syllogistic tasks. By now, several different *models* have been proposed which are trying to reflect and describe human syllogistic reasoning. Each model has its unique theory to solve syllogistic problems. Khemlani and Johnson-Laird, among others, dealt with the challenge to identify the *best* model which is suitable for *every* human reasoner. Within the scope of a meta analysis, 12 of the most prominent models of syllogistic reasoning were evaluated (Khemlani & Johnson-Laird, 2012). Their results show that it is extremely difficult to determine the *best* theory.

One way to rate a model in its syllogistic reasoning is to simply compare the model's responses with the responses given by human reasoners. The more responses of the model correspond to the reasoners' responses, the better the model's *performance* and the more suitable is the model's simulation of human reasoning. Yet, despite being carefully designed, psychological studies must deal with the difficulty that people are highly diverse (regarding education, priming effects, experiences etc.) or, in terms of syllogistic reasoning, pre-conditions may vary (repetition of tasks, feedback etc.). This requires a model to adjust to different circumstances. One possibility for more flexibility is the integration of parameters into the model. A model is thus able to change its reasoning strategy and is therefore capable, albeit limited, to adapt to different reasoners (Riesterer, Brand, & Ragni, 2020a). Consequently, successful parameter configurations may provide information about human reasoning and its underlying concepts.

A well-known cause of a changing strategy in solving syllogistic tasks is giving feedback to the reasoner's responses (Riesterer, Brand, & Ragni, 2020b). This enables the reasoner to use the newly acquired knowledge when giving the next answers. Indeed, reasoners are able to learn and improve their reasoning when feedback is given (Dames, Schiebel, & Ragni, 2020) what requires the model to adapt by changing its parameter configuration. In order to better assess the effects of feedback on a cognitive level and to better understand the changing strategies of the reasoners, it is promising to evaluate the parameter configurations a model performs best with when it is compared to reasoners which were given feedback.

Within the scope of this thesis, three different parameterized models, including TransSet (Brand, Riesterer, & Ragni, 2020), the Probability Heuristics Model (Chater & Oaksford, 1999) and mReasoner (Khemlani & Johnson-Laird, 2013), are selected in order to investigate the changing strategy and mental processes of human reasoning which potentially occur when reasoners receive feedback while solving syllogistic tasks. By changing its parameter configuration each model is capable of adapting to reasoners who received feedback and reasoners to whom feedback has not been provided. Thus, the evaluation of the different parameter configurations can show how reasoners adapt their reasoning based on feedback.

2 Theoretical Background

The best parameter configurations of three selected models (illustrated below) for non-feedback and feedback data are evaluated and compared in order to investigate the underlying mental processes potentially occurring when feedback is given. A parameter configuration of a model contains all parameters a model offers and a specific value assigned to its corresponding parameter. The possible parameter assignments are summarized in Table 2.1 for each model. According to the model, there are numerous configurations which have to be considered. In the following, TransSet (Brand et al., 2020), PHM (Chater & Oaksford, 1999) and mReasoner (Khemlani & Johnson-Laird, 2013) and their possible parameter assignments are described.

2.1 TransSet

TransSet (Brand et al., 2020) tries to simulate human reasoners who are not familiar with solving syllogistic tasks and mainly rely on their intuitive thinking. In particular, it assumes reasoners to focus on term order for instance and to build transitive paths to draw conclusions. Its core is the general concept of transitivity: e.g., if x is smaller than y and y is smaller than z, one can conclude that x is smaller than z. Transferred to syllogistic reasoning, a reasoner may conclude, that *All A are C*, if *All A are B* and *All B are C*. TransSet uses B as its transitive path to make conclusions about A and C. Its algorithm can be divided into two phases: The *direction selection phase* determines if 1) a transitive path can be specified and 2) which direction this path will obtain. The *quantifier selection phase* uses the generated transitive path to infer a conclusion quantifier. It checks the premises for particular (*some* or *some...not*) or negative (*no* or *some...not*) quantifiers, returns NVC if conditions are met or merges the quantifiers to derive a conclusion. To be able to adapt to different individual reasoners, TransSet allows four parameters as shown in Table 2.1: NVC aversion, anchor set, particularity rule and negativity rule. The parameter NVC aversion is a quantifier for the likelihood that TransSet responds NVC. Possible values are *none* (0), *low* (0.5) and *high* (1). It is included in either phase of the algorithm. The higher it is, the less likely it becomes for TransSet returning NVC. The anchor set parameter takes effect within the direction selection phase and defines whether the first or the most-recent term of the syllogistic task is the *anchor point* when determining a direction of the transitive path. The result is a path either shaped A $\rightarrow$ C or C $\rightarrow$ A. The particularity and negativity rule are both binary parameters and toggle specific proceedings within the quantifier selection phase of TransSet. If the particularity rule is active, TransSet derives NVC whenever two particular quantifiers occur in the premises within the generated transitive path. If the negativity rule is active, TransSet will derive NVC if the transitive path is starting with a negative quantifier (Brand et al., 2020). However, the conditions for applying the negativity rule may vary according to TransSet's NVC aversion. If NVC aversion is high for instance, negative quantifiers in *both* premises will be needed in order that TransSet returns NVC.

Table 2.1: Possible parameter configurations of TransSet, PHM and mReasoner. The values for TransSet's NVC aversion are *none* (0), *low* (0.5) and *high* (1). The anchor set can be set to *first* or *most-recent*. The particularity and negativity rule can be activated or deactivated (see Section 2.1 for details). All parameter values for PHM are probabilities and tell how likely it is that PHM makes specific decisions (see Section 2.2 for details). mReasoner's λ reflects the quantity of constructed entities, ε tells hereby the probability whether the constructed entities are complete or not (see Section 2.3 for details). σ and ω specify the probabilities of how mReasoner handles its search for counterexamples and how to proceed if counterexamples were found.

model	parameter	values
TransSet	NVC aversion	{0, 0.5, 1}
	anchor set	{first, most-recent}
	particularity rule	{true, false}
	negativity rule	{true, false}
PHM	p-entailment	[0, 1]
	conf. A	[0, 1]
	conf. I	[0, 1]
	conf. E	[0, 1]
	conf. O	[0, 1]
mReasoner	ε	[0, 1]
	λ	[0, 8]
	ω	[0, 1]
	σ	[0, 1]

2.2 PHM

The Probability Heuristics Model (PHM) relies on simple probability based heuristics (Chater & Oaksford, 1999) and focuses on, just as TransSet, the intuitive thinking of human reasoners. It first generates a set of conclusions by identifying the least informative quantifier occurring in the premises and by deriving alternative conclusions. It then evaluates its own proposed conclusions. Thus, the core of PHM's algorithm can be divided into three *generation* heuristics (G1 - G3), which are used to propose a set of conclusions and two *test* heuristics (T1 - T2) which are rejecting or accepting the proposed conclusions. The main part of the generation heuristics is hereby the so called *min-heuristic* (G1), which defines *"the quantifier of the conclusion to be the same as the quantifier in the least informative premise"* (Chater & Oaksford, 1999). With the second heuristic (G2), also called the *entailment-heuristic* (G2), PHM is able to derive additional alternative conclusions - e.g., instead of concluding *All A are C* it may also be possible and logically valid to conclude *Some A are C*. The *attachment heuristic* (G3) finally specifies the subject of the conclusion by selecting the least *or* the most informative premise, depending on the least informative premise being an end term or not. Both the max-heuristic (T1) and the O-heuristic (T2) decline the proposed conclusions, if they contain uninformative quantifiers. As illustrated in Table 2.1, PHM uses five parameters in order to be able to adapt to different individual reasoners: p-entailment, conf. (confidence) A, conf. I, conf. E and conf. O - all of them are probabilities. The p-entailment parameter defines the probability whether an alternative conclusion is derived or not. Conf. A, conf. I and conf. E specify the probability whether a specific quantifier within a proposed conclusion is accepted or declined by the max-heuristic (T1). Conf. O specifies the probability whether a conclusion, containing the *some...not* quantifier, is accepted or declined by the O-heuristic (T2) respectively. The ranking of quantifiers is hereby defined as $All > Some > No > Some...not$. E.g., if PHM accepts *no* as a quantifier within its proposed conclusion, it will also accept *all* and *some* quantifiers. Trusting a quantifier with a lower rank means that every quantifier with a higher rank is also trusted. Trusting a quantifier with a lower rank and not trusting quantifiers with higher ranks is an invalid configuration for PHM. The more confidence parameters are set to 0, the more likely it is that PHM declines proposed conclusions and returns NVC for an arbitrary task.

2.3 mReasoner

mReasoner (Khemlani & Johnson-Laird, 2013) is based on the Mental Model Theory (MMT) of reasoning (Johnson-Laird, 2010) and builds mental representations of premises by proposing a set of symbolic entities which are assigned to one or more syllogistic terms. Subsequently, first candidate terms to propose conclusions are deduced. mReasoner validates those conclusions by searching for counterexamples. If a counterexample is found, the conclusion is revised, the process restarts or NVC is returned. Just as TransSet, mReasoner allows four parameters in order to adapt to different reasoners: ε, λ, ω and σ. λ specifies the maximum number of initiated symbolic entities and ε tells hereby the probability whether the constructed entities are complete or not. mReasoner is able to derive and propose logically invalid conclusions when using incomplete entities. σ specifies the probability to either search for counterexamples or to immediately return the proposed conclusion. The larger σ is, the more likely mReasoner searches for counterexamples. In case a counterexample is found, the parameter ω specifies the probability to either continue searching for new counterexamples for a now restricted conclusion (e.g. the conclusion *all A are C* becomes *some A are C*) or to return NVC. If it continues to search for new counterexamples and cannot find any further ones, it will return the restricted conclusion. The smaller ω is, the less likely mReasoner starts a new search for counterexamples. In case the search for other counterexamples is interrupted, NVC will be returned.

3 Methods

To examine the effects of feedback on human syllogistic reasoning, three datasets with different pre-conditions were collected in psychological experiments. Each dataset consists of answers to syllogistic problems generated by real people. The quantity of participants varies in each dataset as summarized in Table 3.1. Besides, a few participants aborted the experiment prematurely and corresponding responses are therefore not recorded. In two of the three datasets the reasoners were given feedback on the logical correctness of each of their responses, i.e. if the answers were correct. The participants within the control dataset (data control) have not been given feedback. The participants within the feedback datasets (data 1s and data 10s) have been given feedback for one and 10 seconds respectively.

Table 3.1: Quantity of participants and NVC ratio in each dataset.

dataset	quantity of participants	NVC ratio (%)
data control	39	13.89
data 1s	146	35.36
data 10s	29	36.88

In order to assess the associated cognitive processes potentially occurring when feedback is given to reasoners, the *best* parameter configuration for each model (introduced in Section 2) for each dataset is determined by two different approaches: One approach focuses on the *overall* performance of a model by simply counting the predictions the model made which are equal to the humans' responses within the whole dataset. The other approach is to first focus on every single participant and to determine the most fitting parameter configuration. The most fitting parameters for every participant are finally aggregated and the result can be considered as a parameter configuration which represents most participants. Whereas the first approach provides information about the effects of feedback in general, the second approach allows a more detailed look at each individual within a particular dataset. Since humans are highly diverse and differ greatly in solving syllogistic tasks, it is promising to also consider each individual when investigating syllogistic inferences. In both cases, each parameter and its value within the best configuration for a specific dataset are investigated. One has to mention that PHM and mReasoner do not only propose one conclusion, but are designed to propose several conclusions at once (see Section 2.2 and Section 2.3 for explanation). Since just one response is allowed to be given in the scope of this investigation and comparison, the prediction is selected randomly within the proposed conclusions which may lead to slightly different accuracies when repeating the same calculations.

Due to the great role of NVC responses in reaching high performances (Riesterer, Brand, Dames, & Ragni, 2020) by being the logically correct answer to 37 out of 64 syllogistic problems (58%), and due to its importance regarding performances in feedback data concerning a reduction of NVC aversions through learning effects (Riesterer, Brand, & Ragni, 2020b), the NVC ratio in each dataset (Table 3.1) and each model with its specific configuration was calculated and considered as well. The model's NVC ratio illustrates how many NVC responses a model returns using a specific parameter configuration. Note that the model's NVC ratio with equal parameter configurations should always be equal. Instead of calculating the NVC ratio the model produces when solving the 64 syllogistic tasks, it was calculated by iterating the tasks of a dataset to ensure equal conditions. Since some responses are not recorded, NVC ratios may slightly vary if dataset sizes are different.

Furthermore, the models' best performances were compared to each other and to performances of baselines like the MFA (Most Frequent Answer) model by using the CCOBRA framework[1]. By this, when applied to different datasets, the models' conclusions can be evaluated and ranked in regard to suitability and reliability. The MFA model always returns the most frequent given answer for a task within a dataset. The comparison to the MFA model is an apt option to rate the general performance of syllogistic reasoning models.

Finally, the parameters are evaluated according to their impact on the models' performances in order to evaluate their importance to the model. This may provide information about which parameter to focus on when deriving conclusions about mental processes that potentially occur when giving feedback to reasoners.

Every implementation in the scope of this thesis was coded with Python 3.8.0. The code can be viewed in the GKI repository of the University of Freiburg[2].

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

[2] https://gkigit.informatik.uni-freiburg.de/coco.theses/2020-feedbackparams

3.1 Model Parameter Configuration with Best Performances

The three examined parameterized models are TransSet and python implemented versions of PHM and mReasoner. TransSet is able to take 24 different parameter configurations. In the scope of this thesis, PHM's parameters are still considered to be binary (0 or 1), because e.g. *”even if a reasoner uses p-entailment for 40% of the syllogistic responses, the expected prediction outcome would still be maximized by setting p-entailment to 0”* (Riesterer, Brand, & Ragni, 2020a). Thus, if also considering PHM's quantifier ranking mentioned in Section 2.2, it is configurable with 10 different configurations. The possible parameter values of mReasoner are continuous. To limit the range of possible parameter values of mReasoner, the value sets were defined as follows: For ε, ω and σ the value set was $\{$0.1, 0.2, ..., 1.0$\}$. In case of λ the value set was $\{$0.1, 0,9, 1.7, 2.5, 3.3, 4.0, 4.8, 5.6, 6.4, 7.2, 8.0$\}$. This results in 14641 different configurations of mReasoner. In order to evaluate a model's *performance*, the *accuracy* for each dataset and each parameter configuration was calculated. The accuracy of a model is a simple ratio between correct predictions of a model and all responses within a given dataset.

$$accuracy = \frac{Quantity(Model\ Response = Human\ Response)}{all\ Responses\ within\ the\ Dataset}$$

The higher the accuracy the higher the performance of the model and vice versa. Thus, *accuracy* and *performance* are hereby synonymous terms. By *brute forcing* every parameter configuration of a model, the configuration achieving the highest accuracy within each dataset was identified and documented.

3.2 Model Parameter Configuration Fitting Most Participants

A model's best parameter configuration for a dataset was determined by initially identifying the best configuration for each participant. Therefore, three calculation steps were necessary:

1. The accuracy for each configuration and each participant within a dataset was calculated:

$$accuracy = \frac{Quantity(Model\ Response = Participant\ Response)}{all\ Responses\ of\ the\ Participant\ within\ the\ Dataset}$$

2. The configuration with the best accuracy for each participant was picked. The result of this calculation was a collection structured as shown below:

   ```
   {
     participant1: [{’parameter1=value1|parameter2=value2’...: accuracy1}],
     participant2: [...],
     ...
   }
   ```

   In case of having several configurations with equally large accuracies for one participant, the parameters within the configuration were scored according to the number of different configurations belonging to the participant. E.g., if a participant just represents one single configuration, each parameter within this configuration received a score of one. If a participant represents three configurations with equal accuracies, each parameter and its specific value just received a score of one third.

3. The scores were accumulated by iterating through every participant for each parameter and its value. The result is a collection structured as shown below:

   ```
   {
     ’parameter1=value1’: score1,
     ’parameter2=value2’: score2,
     ...
   }
   ```

Finally, each parameter and its value with the highest score was chosen to determine an overall *best* configuration for the model. This resulting parameter configuration represents most of the participants in a given dataset.

3.3 CCOBRA Benchmark

The CCOBRA framework was used in *prediction* mode to compare the models to each other and to the MFA model. In prediction mode, CCOBRA does not allow models to adjust during prediction phase, only pre-training is enabled. Hence, the models do not adapt their parameter configuration whilst predicting a dataset. They are merely allowed to use the datasets beforehand to optimize their parameters in order to achieve maximum accuracies. This pre-training applies to the MFA model only. Regarding the first approach to evaluate configurations and its performances (see Section 3.1), CCOBRA calculates the accuracy for each model and plots all in one single chart. Regarding the second approach of first focusing on parameter configurations fitting to a single participant (see Section 3.2), CCOBRA is able to illustrate accuracies for individuals by using box plots. In both cases, the performances of the different models can easily be compared and analyzed.

3.4 Parameter with Most Significant Impact on Performances

In order to identify the parameter with most impact on performances a model achieves for a given dataset, the value of one parameter was changed while the other parameters' values were locked within a specific configuration. The accuracies the model achieves with those configurations were compared with each other, the spreads were documented and the average was calculated. A spread is hereby the largest difference between the obtained accuracies. The result is an accuracy spread for each parameter for each dataset.

4 Results

4.1 Model Parameter Configuration with Best Performances

4.1.1 TransSet

Table 4.1 summarizes the parameter configurations of TransSet with best performances within each dataset. The NVC ratio (percentage of NVC responses within all given responses of TransSet) is listed as well. TransSet derives its conclusions by first trying to build a transitive path and second by validating this path (see Section 2.1 for details).

Table 4.1: Parameter configuration of TransSet with best performances and NVC ratio.

dataset	NVC aversion	anchor set	part. rule	neg. rule	accuracy (%)	NVC ratio (%)
data control	1	first	false	true	37.33	25.01
data 1s	0.5	first	true	true	46.88	53.17
data 1s	0.5	most-recent	true	true	46.88	53.17
data 10s	0.5	first	true	true	50.49	53.25
data 10s	0.5	most-recent	true	true	50.49	53.25

Regarding data control, TransSet achieves its best overall performance for reasoners not receiving any feedback with high NVC aversion, if it chooses the first term as the anchor point to build its transitive path and if it does not derive NVC due to two particular quantifiers (*some* or *some...not*) within its generated transitive path. However, TransSet's negativity rule is activated. Due to its high NVC aversion, it will return NVC whenever there are negative quantifiers in *both* premises of its transitive path. This reflects in TransSet's semi-low NVC ratio with the given configuration for data control as shown in Table 4.1. Every fourth response of TransSet is NVC when being configured as described. Regarding data 1s and data 10s, TransSet's NVC aversion shifts from high (1) to low (0.5) and choosing the first or the most-recent term for building its transitive path does not have any effect on the accuracy it achieves within those datasets. However, the particularity rule must be activated in order to achieve high performances in both data 1s and data 10s. A shift from high to low NVC aversion also induces a change regarding the condition concerning TransSet's negativity rule. In this case, it returns NVC if the quantifier of the first premise is negative *and* the quantifier of the second premise of its transitive path is not *all*. Having this configuration, TransSet returns NVC for approximately every second task.

4.1.2 PHM

Table 4.2 shows the parameter configurations of PHM with best performances within each dataset. PHM generates a set of conclusions by identifying the least informative premise and by deriving additional alternative conclusions. It validates its proposed conclusions using test heuristics. If PHM returns several conclusions for one task, a response is randomly selected (see Section 3 for details) in order to calculate performances.

Table 4.2: Parameter configuration of PHM with best performances and NVC ratio.

dataset	p-entailment	conf. A	conf. I	conf. E	conf. O	accuracy (%)	NVC ratio (%)
data control	0	1	1	0	0	35.56	25.01
data 1s	0	1	0	0	0	45.57	56.28
data 10s	0	1	0	0	0	47.67	56.34

To achieve maximum accuracy in data control, PHM accepts *all* and *some* quantifiers within its proposed conclusions. Furthermore, it is *not* deriving alternative conclusions within its entailment heuristic in order to achieve its highest accuracy for data control. Being configured as described, every fourth response of PHM is NVC. Best performances in data 1s and data 10s are achieved by PHM when it only accepts the *all* quantifier within its proposed conclusion. All other conclusions that contain either *some*, *no* or *some...not* as a quantifier are rejected. As in data control, PHM does not derive alternative conclusions for its best performances in data 1s and data 10s. PHM's NVC ratio with its depicted configuration for data 1s and data 10s reaches approximately 50%.

4.1.3 mReasoner

Table 4.3 shows the parameter configurations of mReasoner with best performances within each dataset. mReasoner proposes conclusions by building a set of symbolic entities containing several syllogistic terms and by validating its proposed conclusions. It eventually searches for counterexamples for validation. As PHM, mReasoner may return several conclusions for one task. If so, a prediction is hereby randomly selected to calculate performances.

mReasoner achieves its highest accuracy in data control when there is a probability of 0.2 (see σ in Table 4.3) of searching for counterexamples when the conclusions are proposed. Thus, the probability for returning one of its conclusions without searching for counterexamples is 0.8. If a counterexample is found, it will return NVC with a probability of 0.3 (1 - ω in Table 4.3 regarding data control). The probability for searching for counterexamples after proposing its conclusions shifts from 0.2 (data control) to 1.0 in data 1s and to 0.9 in data 10s as shown in Table 4.3. If a counterexample is found, the chance of immediately returning NVC is 0.7 in data 1s and 0.6 in data 10s (1 - ω in Table 4.3 regarding data 1s and data 10s, respectively). If it starts a new search and cannot find any further counterexamples, it will return the proposed conclusion. If configured as described for data control, there are almost no NVC responses of mReasoner. However, its NVC ratio increases to almost 80% when being configured as described for data 1s.

Table 4.3: Parameter configuration of mReasoner with best performances and NVC ratio.

dataset	ε	λ	ω	σ	accuracy (%)	NVC ratio (%)
data control	0.4	1.7	0.7	0.2	36.33	1.57
data 1s	0.7	4.8	0.3	1.0	43.85	77.44
data 10s	0.9	0.9	0.4	0.9	47.51	69.63

4.1.4 CCOBRA Benchmark Regarding Overall Performance

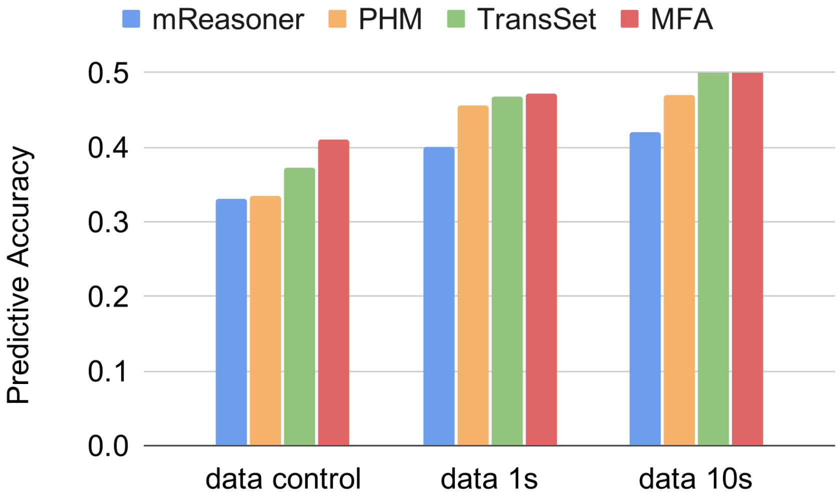


Figure 4.1: Comparison of model performances with CCOBRA in each dataset.

Figure 4.1 compares the models with their parameter configuration achieving the highest accuracy for each dataset. None of the models is able to outperform the MFA model in any case. However, regarding data 1s and data 10s TransSet's performances are almost equal to the performances achieved by the MFA model. PHM and mReasoner are exceeded by TransSet, but their performances are still above 30% in data control and above 40% in data 1s and data 10s. The predictive accuracies in data control, data 1s and data 10s increase constantly. The best performances of every model can be observed in data 10s, followed by the performances reached in data 1s. The least predictive accuracies of all models can be observed in data control with the MFA Model's accuracy reaching approximately 40%.

4.2 Model Parameter Configuration Fitting Most Participants

4.2.1 TransSet

Figure 4.2 shows the configuration of TransSet which fits most participants within data control. In other words, it illustrates how many percent of participants are related to which parameter value. For instance, TransSet describes approximately 75% of the participants within data control when NVC aversion is high (1). Different to the results in Section 4.1.1 for data control, the model fits most participants when its negativity rule is additionally deactivated during the quantifier selection phase. Figure 4.3 shows TransSet's most fitting parameter configuration regarding data 1s. TransSet represents most participants when shifting its NVC aversion from high (1) (Figure 4.2) to low (0.5) (Figure 4.3) with a NVC aversion tendency towards *low*. None (0), low (0.5) and high (1) NVC aversion representations are almost equally distributed among the participants. Furthermore, TransSet needs to activate both its particularity and negativity rule in order to fit most participants within data 1s. Almost the same picture as in Figure 4.3 can be seen regarding TransSet's participant parameter representations for data 10s (Figure 4.4). However, less participants can be assigned to a high (1) NVC aversion.

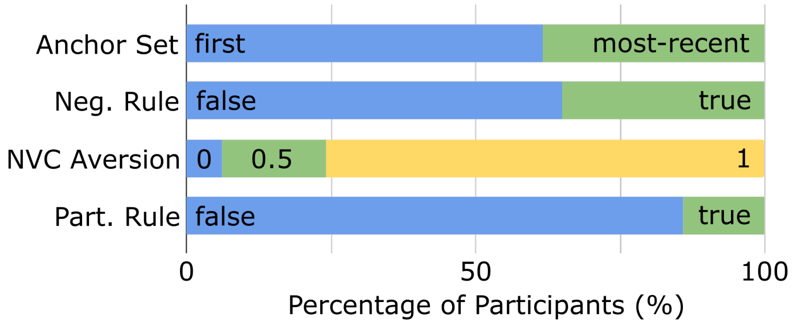


Figure 4.2: Parameter configuration distribution of TransSet for data control.

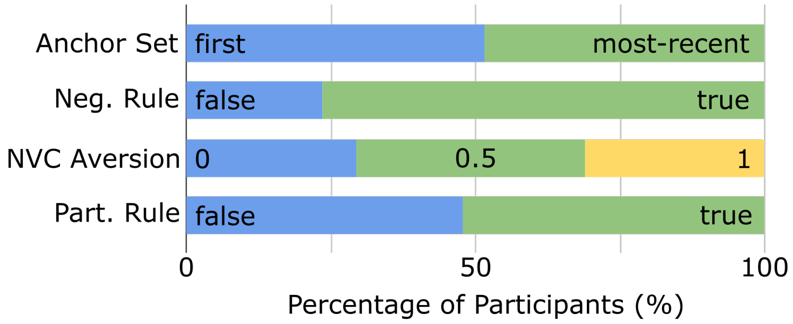


Figure 4.3: Parameter configuration distribution of TransSet for data 1s.

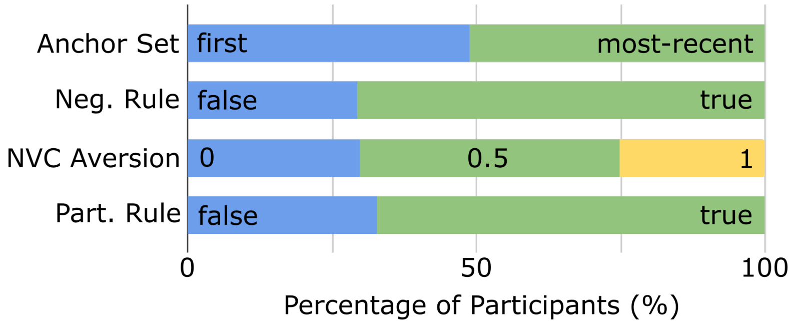


Figure 4.4: Parameter configuration distribution of TransSet for data 10s.

Table 4.4 summarizes the results for the parameter configurations of TransSet fitting most participants for each dataset and shows the NVC ratio the model produces when being configured as illustrated. TransSet returns no single NVC response, when its NVC aversion is high (1), when it chooses the first term to construct its transitive path and when its particularity and negativity rule within the quantifier selection phase are both deactivated.

Table 4.4: Parameter configurations of TransSet fitting most participants and NVC ratio.

dataset	NVC aversion	anchor set	part. rule	neg. rule	NVC ratio (%)
data control	1	first	false	false	0.0
data 1s	0.5	first	true	true	53.17
data 10s	0.5	most-recent	true	true	53.25

4.2.2 PHM

Figure 4.5 shows which parameter values of PHM stands the majority of participants in data control. It describes most participants when its max-heuristic (T1) accepts *all* and *some* quantifiers within its conclusions. There are slightly more participants within data control who are represented by PHM if it accepts the *no* quantifier as well. The value distribution for PHM's confidence O shows a small advantage for 0. Thus, most participants are represented by PHM if it declines *some...not* quantifiers. Figure 4.6 and Figure 4.7 show PHM's parameter distribution within data 1s and data 10s. It fits most participants if just accepting *all* quantifiers within its conclusions for both datasets. Regarding data 10s, the parameter value distribution is pretty one-sided. Every percentage of the dominating parameter value reaches at least 75%.

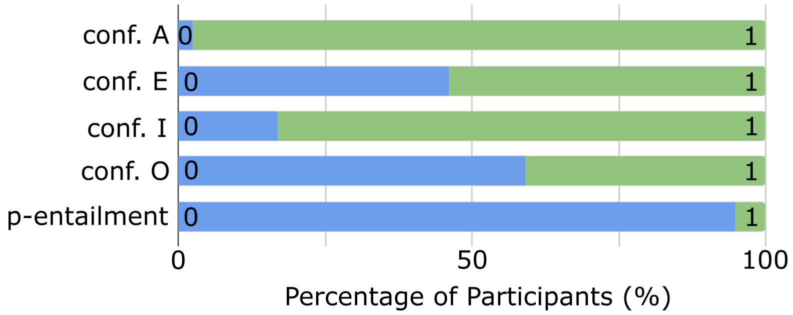


Figure 4.5: Parameter configuration distribution of PHM for data control.

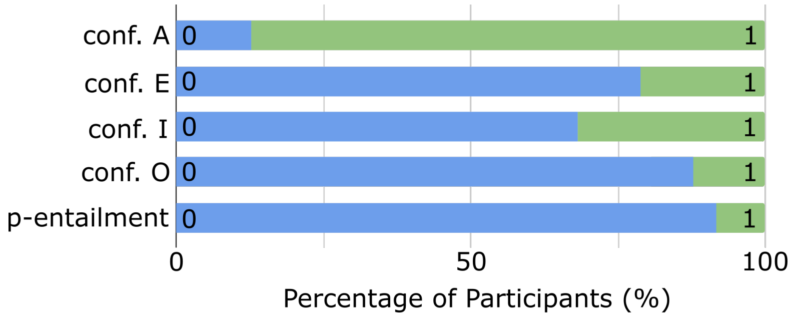


Figure 4.6: Parameter configuration distribution of PHM for data 1s.

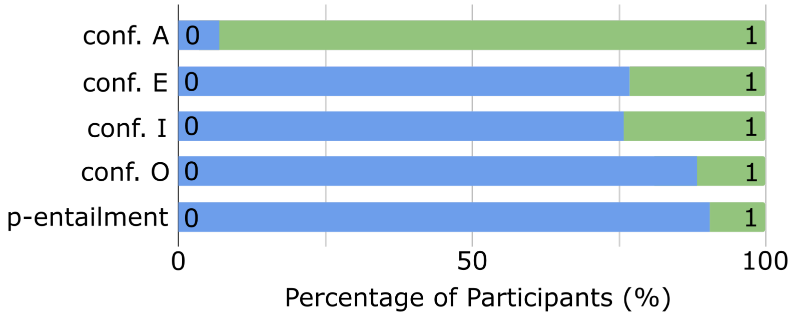


Figure 4.7: Parameter configuration distribution of PHM for data 10s.

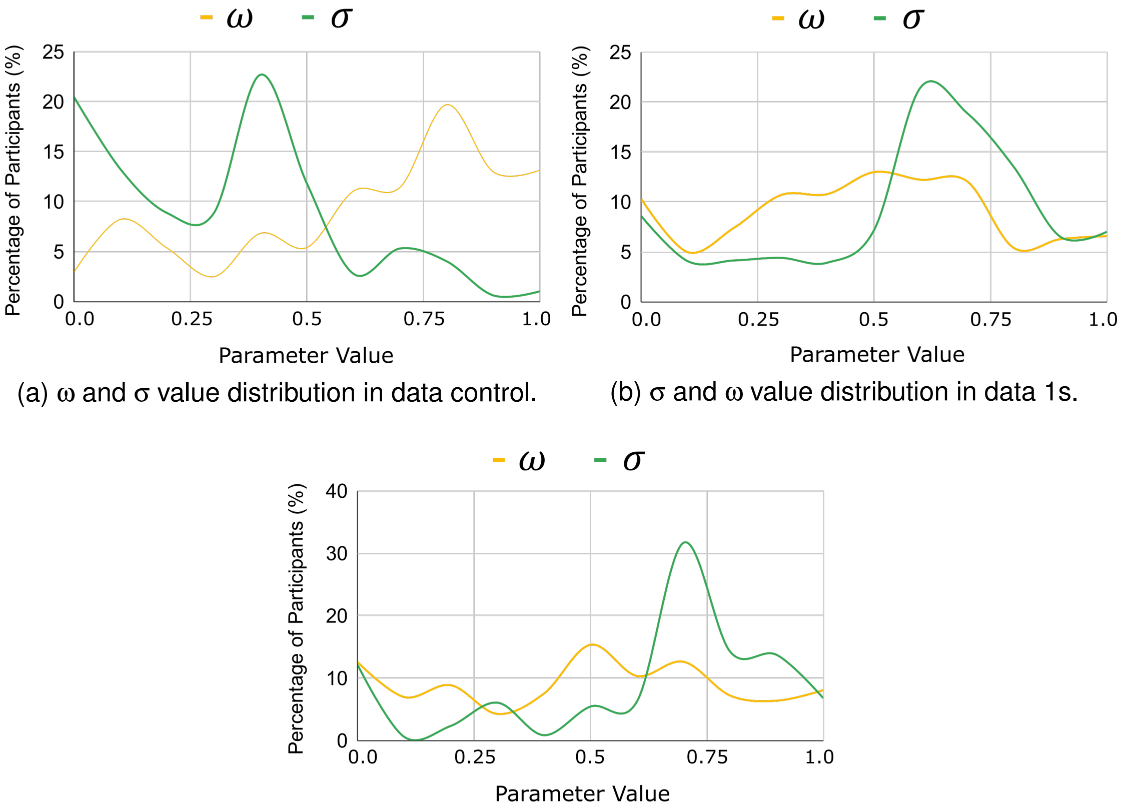
Table 4.5 summarizes the results shown in Figure 4.5, Figure 4.6 and Figure 4.7 and outlines the NVC ratio PHM produces when configured as specified. The configurations PHM fits most participants with are the same as in Table 4.2 regarding overall performances in data 1s and data 10s. In data control however, PHM additionally accepts the *no* quantifier besides accepting *all* and *some* quantifiers within its proposed conclusions. Not rejecting *all*, *some* and *no* quantifiers leads to hardly returning NVC. Just 6.28% of the responses of PHM are NVC if configured as described for data control.

Table 4.5: Parameter configurations of PHM fitting most participants and NVC ratio.

dataset	p-entailment	conf. A	conf. I	conf. E	conf. O	NVC ratio
data control	0	1	1	1	0	6.28
data 1s	0	1	0	0	0	56.28
data 10s	0	1	0	0	0	56.34

4.2.3 mReasoner

The parameters of mReasoner are continuous. Thus, a line chart is chosen to roughly approximate and describe its parameters. Since the line charts for ε and λ did not show great volatile behaviour, just the line charts for ω and σ are shown in Figure 4.8. Nonetheless, the best values for ε and λ in the scope of this investigation are shown in Table 4.6. Figure 4.8a shows ω's and σ's parameter assignments fitting most participants in data control. mReasoner describes most of the participants in data control if there is just a probability of 0.4 of searching for counterexamples. On the other hand, the probability of mReasoner's ω parameter is very high in this matter. If mReasoner searches for counterexamples and a counterexample is found, it is searching for new counterexamples with a probability of 0.8 when representing most participants. Thus, the probability for immediately returning NVC is just 0.2. The values for ω and σ which describe most participants in data 1s and data 10s are shifting towards 0 and 1, respectively. The probability for searching for counterexamples right after the conclusion is proposed (σ) shifts from 0.4 (data control) to 0.6 (data 1s) and 0.7 (data 10s). The probability for searching for new counterexamples after a counterexample is found (ω) decreases from 0.8 (data control) to 0.5 (data 1s and data 10s).



(a) ω and σ value distribution in data control.

(b) σ and ω value distribution in data 1s.

(c) σ and ω value distribution in data 10s.

Figure 4.8: Parameter distribution of mReasoner.

Table 4.6 summarizes the results for the parameter configurations of mReasoner fitting most participants for each dataset. Additionally, it shows the NVC ratio mReasoner produces when configured as illustrated. The NVC ratio of mReasoner and its configuration for data control is the lowest. Just 15.87% of its responses are NVC. However, the NVC ratio increases to more than approximately 50% and 60% when configuring mReasoner as shown in Table 4.6 for data 1s and data 10s, respectively.

Table 4.6: Parameter configurations of mReasoner fitting most participants and NVC ratio.

dataset	ε	λ	ω	σ	NVC ratio (%)
data control	0.2	0.1	0.8	0.4	15.87
data 1s	0.0	8.0	0.5	0.6	53.62
data 10s	0.2	0.1	0.5	0.7	65.67

4.2.4 CCOBRA Benchmark Regarding Individual Performance

Figure 4.9, Figure 4.10 and Figure 4.11 illustrate accuracies for individuals. One dot represents the percentage of equal responses between the model and the corresponding individual. The parameter configuration of the models were chosen according to the results shown in Table 4.4, Table 4.5 and Table 4.6 in respect to each dataset. TransSet once again outperforms PHM and mReasoner in every dataset. However, just as shown in Figure 4.1, TransSet is not able to outperform the MFA model. The predictive accuracies for data control and data 1s are approximately equally distributed whereas the accuracies for data 10s are clustering inside upper and lower value areas.

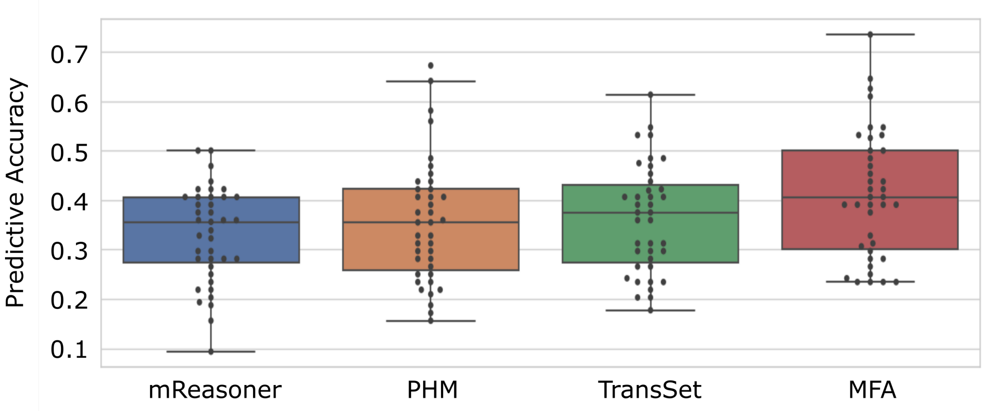


Figure 4.9: Boxplots with comparison of accuracies for individuals with CCOBRA regarding data control.

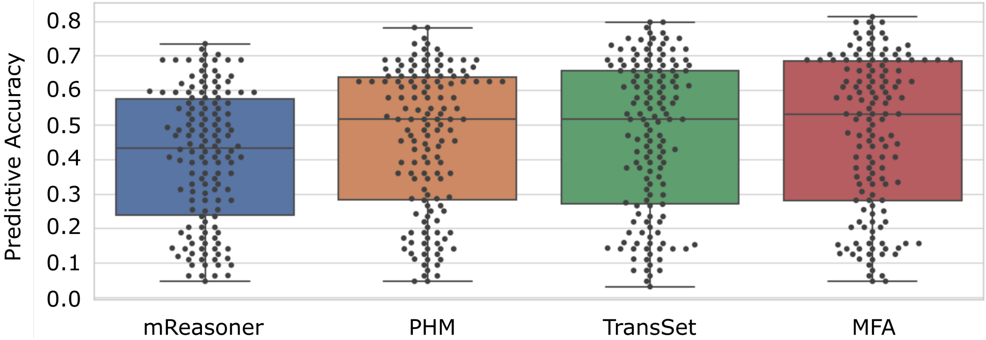


Figure 4.10: Boxplots with comparison of accuracies for individuals with CCOBRA regarding data 1s.

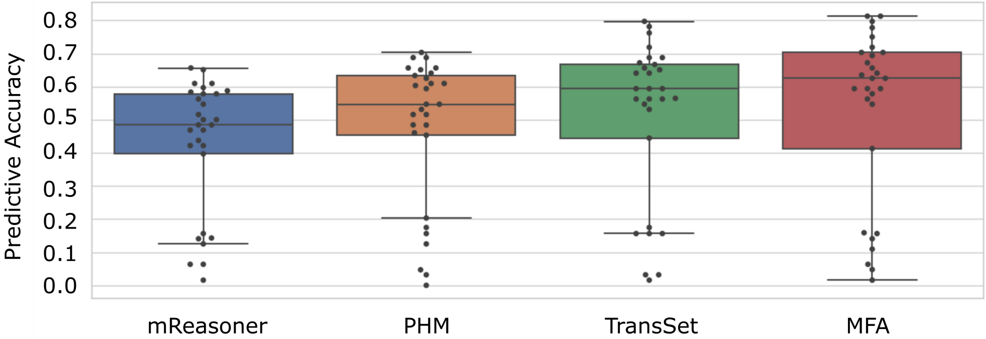


Figure 4.11: Boxplots with comparison of accuracies for individuals with CCOBRA regarding data 10s.

4.3 Parameter with Most Significant Impact on Performances

Figure 4.12, Figure 4.13 and Figure 4.14 shows the parameter impact on the model's performances for each dataset. The larger the average accuracy spread, the bigger the parameter's impact on the model's performances for a specific dataset. TransSet's negativity rule and its NVC aversion cause the biggest average spreads when examining data 1s and data 10s. Changing the anchor point (first or most-recent) does not lead to big differences in performances of TransSet in data control and there are no differences at all regarding performances in data 1s and data 10s. As illustrated in Figure 4.13, PHM's p-entailment parameter contributes most to the performance in data control. In data 1s and data 10s almost every parameter (except confidence O) influences a shift of PHM's average performance by approximately 7%. Changing PHM's acceptance of *some...not* quantifiers (confidence O) within the proposed conclusion hardly influences its average performance. Regarding Figure 4.14, changing the probability to either search for counterexamples or to immediately return the proposed conclusion (σ) dramatically affects mReasoners performances in any dataset. The other parameters of mReasoner are not able to change the model's performance by more than 4% on average.

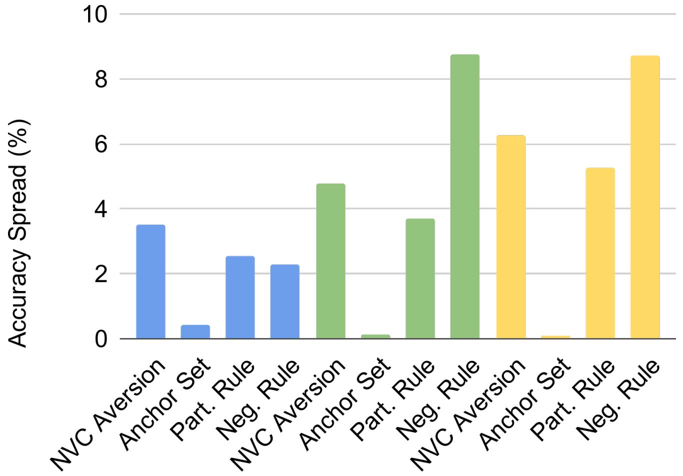


Figure 4.12: TransSet's accuracy spreads regarding each parameter. Blue: Data control. Green: Data 1s. Yellow: Data 10s.

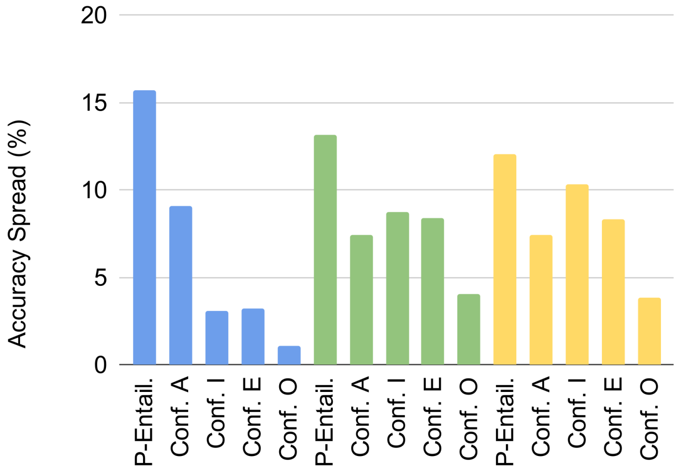


Figure 4.13: PHM's accuracy spreads regarding each parameter. Blue: Data control. Green: Data 1s. Yellow: Data 10s.

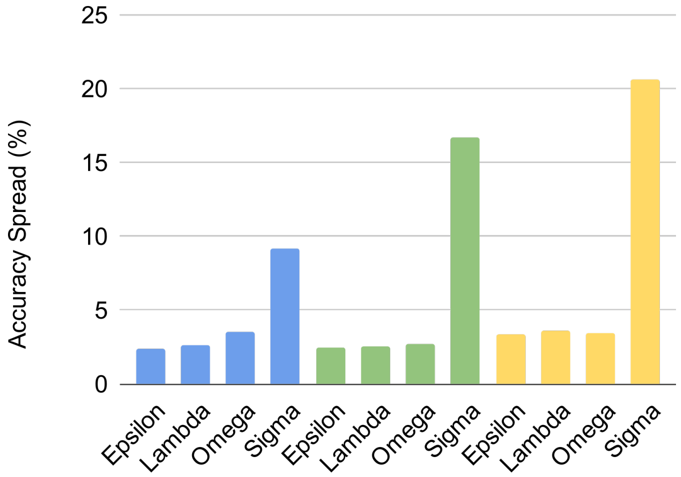


Figure 4.14: mReasoner's accuracy spreads regarding each parameter. Blue: Data control. Green: Data 1s. Yellow: Data 10s.

5 Discussion

The human mind has always been a subject of interest and speculation; who would not want to know what the other one thinks or is about to do? Within the last couple of decades, many psychological experiments have shown that most people seem to think, judge and make decisions according to certain ”rules” (Kahneman, 2017). In order to appropriately discuss the results presented within this thesis, it is important to be aware that human thinking works by the ”law of least effort” (Kahneman, 2017). It means that usually, people spontaneously answer with an intuitive solution which emerges quickly and with ease. If the search for one fails when faced with a difficult problem (such as syllogisms), they switch to a slower, more deliberate and effortful form of thinking. With this in mind, some observations made in this study can be interpreted.

In the here used query where participants were asked to respond to a set of syllogisms, the influence of intuitive thinking can be clearly seen. Without any feedback, people tend to have high NVC aversions when answering syllogistic questions. This can be explained by the unease that many people experience when they are unable to give a ”satisfying” answer; responding NVC feels unfamiliar and wrong, and is hence, avoided under all circumstances (Ragni, Dames, Brand, & Riesterer, 2019). In fact, NVC is the logically correct answer to most syllogistic tasks: 58% of syllogistic problems are solved correctly when responding with NVC. This answering behaviour of the participants can be influenced by feedback. As seen in Table 3.1, the rate by which NVC answers are given, increases significantly in feedback datasets. Remarkably, this is not in congruence with giving more correct answers (Riesterer, Brand, & Ragni, 2020b). Indeed, people seem to feel less alienated from NVC and answer NVC more confidently or, in other words, included this option into their intuitive thinking. To put it bluntly, the participants benefit from the proportionally greater abundance of NVC (Dames et al., 2020) and still, refrain from a more demanding cognitive operation required when solving logical puzzles. These observations can be reproduced by certain computational approaches. In the following the three different parameterized models (TransSet, PHM and mReasoner) which aim to represent human reasoning, when challenged with syllogistic questions, will be discussed.

In the picture that emerges from the evaluation of results, the NVC ratio can be considered the most reliable factor to interpret changes in participants' answering upon feedback, not the correctness of answers itself. This will be discussed in more detail later on. Albeit, in many computational models, the NVC ratio is often neglected (Riesterer, Brand, Dames, & Ragni, 2020).

To begin with, the parameter settings for each model will be described for the case when the model represents the participants' reasoning behaviour best. Next, the impact of those parameters which are responsible for the best performance of these models will be evaluated. Eventually, the models will be compared to the MFA baseline and the findings will be discussed.

5.1 TransSet

TransSet represents reasoners to whom feedback has not been provided (data control) best when it is configured as follows: the NVC aversion is set to high (1) and its particularity rule is disabled. Moreover, regarding the overall performance, i.e. the ratio between correct predictions of a model and all responses within a given dataset, the negativity rule is enabled (Table 4.1). In contrast, regarding the participant fitting , i.e. the model's parameter configuration fitting most participants, it is inactivated (Table 4.4). This leads to TransSet's low NVC rate of 25.01% in Table 4.1 and to 0.0% in Table 4.4.

High NVC aversion regarding data control suggests that reasoners, without having received feedback, tend to avoid NVC responses. Indeed, 13.89% of the answers given by the participants in data control (Table 3.1) are NVC. This answering behaviour of the participants can be influenced by feedback towards higher NVC answering rates (35.36% in data 1s and 36.88% in data 10s). Correspondingly, TransSet shifts its NVC aversion from high to low (Table 4.1, Table 4.4), and both the particularity and the negativity rule are activated (Table 4.4). By this, two things can be achieved: first, the model's accuracy increases (overall performance). Second, the individual performance is reproduced (participant fitting). Hence, TransSet describes the reasoners within feedback datasets as *less reluctant* to responding NVC - their NVC aversion decreases.

Parameter impact investigations for TransSet in feedback datasets show that the NVC aversion parameter and the negativity rule influence the model's performance greatest (Figure 4.12). With the finding by Riesterer & Brand et al. (2020b) in mind that despite feedback participants not necessarily answer more correctly, the importance of those two parameters indicates two psychological processes. The lower aversion to responding NVC can be explained by a priming effect, which means that once the idea that answering NVC is often correct is activated, people are more likely to think of NVC and consequently, tend to answer NVC more often.

The negativity rule checks the premises for *no* or *some...not* and consequently, immediately returns NVC, when activated. Else, TransSet will continue searching for the correct answer. Consider the following syllogistic example: No diver (A) is an adventurer (B) and all adventurers (B) are treasure hunters (C). Here, no conclusion between A and C can be derived and NVC is the correct answer. With active negativity rule the transitive path is interrupted due to the negative quantifier and NVC promptly returned. Hence, the negativity rule can be interpreted as an additional thinking step in case it is inactivated because the search for a correct answer will not be cut short by *no* or *some...not*. This requires more effort and thus, is often avoided by the average participant. However, when the negativity rule is active, it is possible that either a learning effect or “laziness” can be observed. Participants may become skilled in the task of identifying *no* or *some...not* in the premises as indicators of NVC. Or, as described above, use their newly acquired knowledge from feedback that in most cases NVC would be the correct answer and respond NVC more generously without laboriously thinking about the answer.

Remarkably, there are no significant differences between 1s feedback data and 10s feedback data concerning the best parameter configuration and, as will be shown for PHM and mReasoner, too. The potential conclusions will be discussed towards the end of the discussion.

5.2 PHM

For data control the Probability Heuristics Model (PHM) achieves its best performance regarding the overall performance when the *all* and the *some* quantifiers within its proposed conclusions are accepted and p-entailment disabled (Table 4.2). Concerning participant fitting, the best performance can be seen when additionally, the *no* quantifier is accepted with p-entailment still being disabled (Table 4.5). Yet, these findings have to be interpreted with care. In fact, there is little difference regarding accuracy outcome in data control, no matter if two, three or four out of the four quantifiers (conf. A, I, E, O) are accepted[1]. This can be explained because of the random selection of a single solution from a set of possible answers provided by PHM (see Section 2.2) and the resulting differences in accuracies for same calculations.

Accepting a quantifier leads during the test heuristics (T1, T2) to a conclusion in accordance with the quantifier, which will be returned instead of NVC. Thus, the more quantifiers are trusted, the lower PHM's NVC ratio. This is well represented for data control where the NVC ratio is 25.01% for overall performance (Table 4.2) and 6.28% for participant fitting (Table 4.5). Concerning participants, the confidence in those quantifiers might illustrate how people *trust* their conclusions and hold on to their decisions.

When feedback is given, PHM achieves the best performance when only the A quantifier is trusted, both for overall performance (Table 4.2) as for participant fitting (Table 4.5). This is interesting because it indicates that participants tend to become more insecure or sceptical in their decision to trust conclusions represented by the quantifiers, and rather reply NVC. In this study the NVC response rate over time by participants who received feedback, was not analysed. Yet, Ragni & Dames et al. (2019) show that *"the likelihood to respond NVC increases over the time-course of the experiment".* Having had the experience that their conclusions as represented by the quantifiers were often incorrect, people may develop doubts and the higher NVC ratio can be considered to signify growing distrust. Similar to TransSet, PHM shows no significant differences regarding best parameter configurations in the two feedback datasets (1s and 10s).

[1] https://gkigit.informatik.uni-freiburg.de/coco.theses/2020-feedbackparams/-/blob/master/thesis_results/phm_results.csv

As shown in Figure 4.13, in case p-entailment is disabled, the overall performances increase dramatically both for data control and feedback data[2]. When p-entailment is active, alternative solutions are searched for. Transferred to human reasoning, this can be interpreted as the difficulty that a human brain is challenged with when it is trying to juggle abstract lines of reasoning simultaneously.

5.3 mReasoner

For mReasoner which contains four parameters, only two parameters (σ and ω) show considerable differences between the datasets (control vs. feedback datasets, Table 4.3 and Table 4.6). σ and ω also account best for the change in human reasoning upon feedback as will be shown in the following. For this reason, only σ and ω will be discussed in detail.

mReasoner simulates participants from data control best when ω's probability is high and σ's probability is low, both for overall (Table 4.3) as for individual performance (Table 4.6). When a parameter's probability is high, counterexamples for conclusions proposed by mReasoner are searched for. σ is superordinate from which follows that when σ is not looking for counterexamples (low probability), the proposed conclusion(s) by mReasoner will be returned. However, if it searches for counterexamples and a counterexample is found, ω becomes activated. The lower its probability, the more likely NVC will be replied. In case of high probability a new search for counterexamples is initiated (concerning a now restricted conclusion, see Section 2.3) and if none is found, the (restricted) conclusion will be returned. For the configuration representing data control, it reads as follows: the conclusion is returned, and for the unlikely event that a counterexample is searched for and one is found, ω continues the search for further counterexamples. In whatever case, NVC is seldomly returned. The analysis of data control indicates that the participants were people not trained in solving syllogisms because they very likely made the first conclusion that came to mind and moreover, were not familiar with NVC.

As can be seen in Table 4.3 and Table 4.6, the probability distribution for σ and ω is reversed upon feedback. σ possesses a high probability to search for counterexamples whereas ω's is low. This is to be equated with an immediate search for counterexamples and once one is found, NVC is responded. Related to human reasoning, it can be

[2] https://gkigit.informatik.uni-freiburg.de/coco.theses/2020-feedbackparams/-/blob/master/thesis_results/phm_results.csv

construed as enhanced scepticism from the participants. Having learned from feedback that not only their answers were partly wrong but also that in most cases NVC proofs to be correct, potentially, they look for counterexamples for a conclusion which they would have made if no feedback was given. In case the participants find one, NVC is responded despite the fact that a search for further counterexamples might render the correct answer (if it is not NVC). In common with PHM, the parameter configuration indicates either scepticism or a penchant for effortless thinking represented by the acceptance of the next best solution.

Again, as TransSet and PHM, mReasoner does not display significant differences regarding the best configurations for σ and ω in feedback datasets.

5.4 1s vs. 10s Feedback

Interestingly, no significant difference in regard to the parameter configurations between the 1s feedback dataset and the 10s feedback dataset can be found in any of the models. This might be an indication for the difficulty that people are facing when trying to embrace the concept of logical thinking required for solving syllogisms. It might also hint to the power that intuitive thinking possesses over participants: regardless of the time allowed to evaluate the feedback information, people readily respond NVC, once the idea of NVC most likely being the correct answer has been introduced.

Remarkably, for individual performances the predictive accuracy values by the models are spread in two clusters (low vs. high accuracy) regarding data 10s as shown in Figure 4.11, whereas the accuracies are equally distributed in the other datasets (Figure 4.9, Figure 4.10). This holds true for any of the analysed models. In case of 1s feedback data with clustering found at high accuracy levels, it can be assumed that the models accurately predict participants' behaviour. Yet, for 10s feedback data, this is only partially correct anymore: some participants are insufficiently described as represented by low accuracy levels. This might imply a learning effect which some participants experience and which comes into effect after a certain time period. Provided those participants actually understand syllogistic reasoning after receiving feedback and thus, give more logially correct responses, the accuracy by the models is decreased.

5.5 CCOBRA

CCOBRA assesses the validity of each model through comparison to the baseline Most Frequent Answer (MFA) model (Figure 4.1, Figure 4.9, Figure 4.10, Figure 4.11). Here, it can be seen that TransSet's performance in any of the datasets (data control and feedback datasets) and in any of the analysis (overall and individual performance) outperforms PHM as well as mReasoner. Moreover, TransSet's performance is close to the one of MFA which implies a certain reliability of TransSet to derive conclusions about mental processes upon feedback. According to TransSet participants become less reluctant to responding NVC when feedback is given as could be shown in this thesis.

In this study, three different computational approaches for simulating human reasoning were evaluated. Yet, to fully understand the human mind and to be able to correctly reproduce it by models, many more psychological, biochemical and neurological analyses and computational investigations must be carried out.

References

Brand, D., Riesterer, N., & Ragni, M. (2020). Extending TransSet: An Individualized Model for Human Syllogistic Reasoning. *Proceedings of the 18th International Conference on Cognitive Modelling (ICCM 2020)*, 17-22.

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

Dames, H., Schiebel, C., & Ragni, M. (2020). The Role of Feedback and Post-Error Adaptations in Reasoning. *Retrieved from osf. io/dy3gr*.

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

Kahneman. (2017). *Thinking, fast and slow.*

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

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

Ragni, M., Dames, H., Brand, D., & Riesterer, N. (2019). When does a reasoner respond: Nothing follows? In *Cogsci* (pp. 2640–2546).

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 *Proceedings of the 42nd annual conference of the cognitive science society.*

Riesterer, N., Brand, D., & Ragni, M. (2020b). Feedback Influences Syllogistic Strategy: An Analysis based on Joint Nonnegative Matrix Factorization. *Proceedings of the 18th International Conference on Cognitive Modelling (ICCM 2020)*, 223-228.

Störring, G. W. (1908). Experimentelle Untersuchungen über einfache Schlussfolgerungsprozesse. *Archiv für die gesamte Psychologie*, *11*, 1-127.

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *science*, *185*(4157), 1124–1131.

Woodworth, R. S., & Sells, S. B. (1935). An atmosphere effect in formal syllogistic reasoning. *Journal of experimental psychology*, *18*(4), 451.

Declaration

I declare that I have produced this Bachelor's thesis entitled 'Evaluation of Model Parameter Configuration Based on Feedback Data' independently and without improper external assistance. I have identified all word-for-word quotations of other authors, as well as comments based closely on other authors' ideas, and I have listed the relevant sources.

Engelhardt

Nils Engelhardt, Freiburg, February 1, 2021