Reverse-Engineering of Boolean Concepts: A Benchmark Analysis

Felix Kettner (felix.kettner@hsw.tu-chemnitz.de)
Elisa-Maria Heinrich (elisa-maria.heinrich@hsw.tu-chemnitz.de)
Daniel Brand (daniel.brand@cognition.uni-freiburg.de)
Marco Ragni (marco.ragni@hsw.tu-chemnitz.de)
Department Behavioural and Social Sciences, Technische Universität Chemnitz,
Straße der Nationen 62, 09111 Chemnitz

Abstract

For a long time the human capability to form hypotheses from observations has been in the focus of research in psychology and cognitive science. An interesting case is to form hypotheses about the underlying mechanisms of technical systems. This process is called reverse-engineering, i.e., to identify *how* a system works. Research so far has focused on identifying general principles of the underlying reasoning process and lead to the development of at least three general approaches. This paper investigates the predictive power of existing models for each individual reasoner for the first time, i.e., can the individual reasoner reverse engineer the Boolean Concepts from observations. Towards this goal, we (i) defined a modeling task on the individual level, (ii) adapt or re-implement existing models for Boolean Concept learning to make predictions on the individual level, (iii) identify base-line models and additional strategies, and (iv) evaluate the models. By focusing on the individual level, we uncover limitations of current state of the art and discuss possible solutions.

Keywords: Boolean concepts; mental models; algebraic complexity; minimal description; reverse-engineering; benchmark

Introduction

Imagine a living room with a single lightsource in the middle of the room, and several doors with lightswitches next to each door. The basic assumption is that every single switch is included into the circuit and therefore has an influence on the condition of the light. Given this, and the fact that every switch can have two different states, i.e., *on* and *off*, there are several combinations of these states which will result in a shining lightbulb, and the remaining possible combinations will turn the light *off*. This concept can be reduced and depicted as shown in Figure 1 by utilising just a representation of the switches and the lightsource. If you had the task to figure out and describe the valid combinations of switches to light the bulb, how would you proceed? Presumably, you would try different combinations and finally come up with an corresponding answer. Such an answer could look like "*Switch a* has to be turned *on* and *switch b* has to be turned *off* to turn *on* the light.". By answering in such a way we intuitively tend to use so called "Boolean concepts" to develop an idea of the underlying electric circuit. Boolean refers to the fact that a variable, in the example above the single lightswitches and the lightbulb, can only have two different states: *on* or *off*. In logical circuits they are represented by *true* and *false*. Furthermore Boolean operators are a way to combine variables or states with other ones in a logical way to describe conditions for a certain target state.

Two basic Boolean operators are *AND* and *OR* which are used to combine variables just like in the given example above: "*Switch a* has to be on *AND switch b* has to be off to turn on the light.". Another operator is *NOT*, which reverses the state of a variable (e.g., the state "off" could also be desribed as "*NOT* on"). Instead of writing *AND*, *OR* and *NOT* in Boolean algebra the symbols $\wedge$, $\vee$ and $\neg$, respectively, are used. For simplicity, we also refer to the switches only with a, b and c, respectively. In our example this would lead to the expression $a \wedge \neg b$ to describe when the light turns on.

Although the basic elements of Boolean concepts are quite simple, the combination of several variables can become very complex and therefore hard to comprehend for humans. The effect of increasing complexity leading to more difficulties for humans to understand such expressions is known as the Shephard trend based on work of Shepard, Hovland, and Jenkins (1961) and confirmed by various other authors (e.g., Smith, Minda, & Washburn, 2004; Love, 2002; Feldman, 2000). Since the inception of this trend a lot of attempts have been made to find a suitable measurement for the complexity of Boolean Concepts to predict human performance in this field accordingly. Some of the most prominent theories are Minimal Descriptions (Feldman, 2000), Algebraic Complexity (Feldman, 2006) and Mental Models by Goodwin and Johnson-Laird (2011).

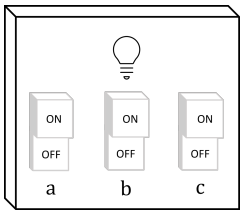


Figure 1: Example of three switches and a lightbulb

While those previous approaches focused on modeling the statistical aggregate of all participant's responses, the focus of this paper is to identify the reasoning difficulty for each individual, i.e., when does the task become too difficult to solve correctly? Hence, we implemented the mentioned theories with a mechanism to adapt to an individual reasoner, compared their accuracy when accounting for the correctness of individual participants and investigated possible extensions.

Boolean concepts

We briefly introduce some necessary background on Boolean concepts. The first step in understanding Boolean concepts is to grasp Boolean variables. Boolean concepts are build upon variables, which can only have the two distinct states of *true* or *false* respectively, when talking about circuits, *on* and *off*. Based on this we can already depict a simple circuit as shown in Figure 2 where the state of the switch equals the state of the whole system, i.e., when the switch is *on*, the light will be *on*.

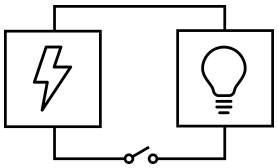
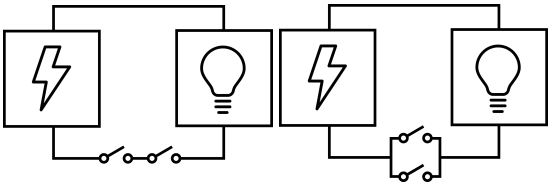


Figure 2: Depiction of a simple circuit where the state of the switch equals the state of the light.

But when adding more switches to the system we need operators to describe in which way these switches depend on each other and impact the state of the whole system. Figure 3 gives an example of two possible configurations of a circuit with two switches which now leads to the basic Boolean operations: The conjunction (with the operator *AND*; $\wedge$) and the disjunction (with the operator *OR*; $\vee$).



(a) Conjunction (*AND*; $\wedge$) (b) Disjunction (*OR*; $\vee$)

Figure 3: Depiction of circuits.

Conjunctions are evaluated to be fulfilled, hence true, if all the combined single statements are fulfilled. A disjunction is fulfilled if at least one of the combined statements is fulfilled. Therefore, refering to Figure 3a, the circuit shows the conjunction concept where both switches (i.e., $a \wedge b$) have to be on in order to turn the light on. The circuit in Figure 3b shows the disjunction concept where it is sufficient that solely one switch is on in order to turn the light on, but still both switches on will also lead to a shining lightbulb. The third basic operation of Boolean concepts is the negation, which serves to reverse the state of a variable or statement (*NOT*; $\neg$).

One peculiarity about Boolean concepts is, that although the basics are quite simple, the combination of several variables can easily get very complex. With three variables already eight combinations are available as shown in Figure 4 with the Boolean concept $(a \vee c) \wedge \neg b$ used as an example.

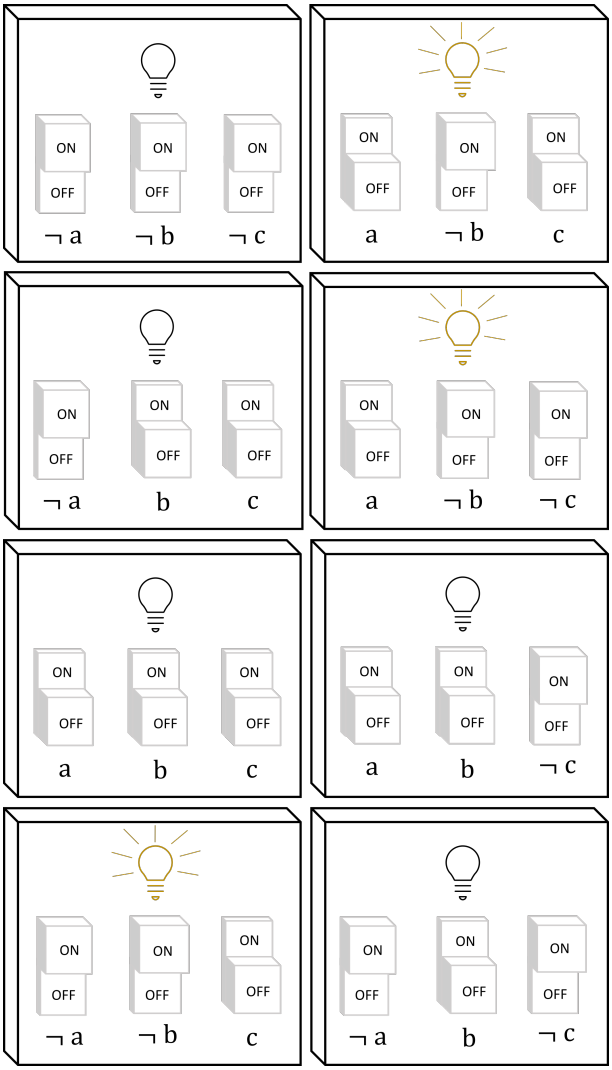


Figure 4: All possible combinations for a Boolean concept with three variables representated as switches. The instances for the concept $(a \vee c) \wedge \neg b$ are highlighted with an active lightbulb.

Approaches for Estimating Difficulty

In this section we introduce the approaches that we use in our analysis as an estimate for the difficulty of Boolean concepts (operationalized by the correctness when solved by participants). For the scope of this paper, we selected only approaches where either a full implementation was available or the respective difficulty estimates were reported by Goodwin and Johnson-Laird (2011). This ensures that the results are comparable and no discrepancies due to a different understanding of the approaches occur.

Minimal Description

For each Boolean expression exists a minimal description length. For example, $(a \wedge b) \vee (a \wedge \neg b) \vee (\neg a \wedge b)$ can be shortened to the minimal description $a \vee b$ which can not be shortened any further. Due to the fact that deriving such

minimal descriptions from complex Boolean expressions is not computationally tractable, Feldman (2000) used a set of heuristics to find the corresponding minimal descriptions for the Boolean concepts used in the given dataset. The minimal description value then equals the amount of used variables in the respective minimal description as shown in the examples in Table 1. Based on the Shepard trend (Shepard et al., 1961)

Table 1: Examples for minimal description values

Minimal description	Value
$a \wedge \neg\, b$	2
$(a \wedge \neg\, b) \vee (a \wedge c)$	4
$\neg(((a \wedge \neg\, b) \wedge c) \vee ((\neg\, a \wedge b) \wedge \neg\, c)))$	6

and Feldman (2000) this classification of Boolean concepts should be able to predict their difficulty.

Algebraic Complexity

The approach of algebraic complexity by Feldman (2006) is based on a decomposition of Boolean expressions into underlying regularities instead of using the minimal description length. Therefore Boolean Concepts are decomposed to their most basic levels, which are single variables on the one hand and the concepts that are combining two variables on the other hand. These atomic elements are the building blocks of each Boolean expression. By analysing the complexity of combinations of those elements within a Boolean expression Feldman (2006) calculates the Algebraic Complexity value. Following Feldman (2006), this principle should perform better in predicting the difficulty of Boolean concepts than minimal description length due to the assumption, that humans are trying to identify statistical regularities in data sets. The corresponding values of Algebraic Complexity for Boolean concepts presented in this paper are taken from Goodwin and Johnson-Laird (2011) who calculated them based on a Matlab Suite provided by Jacob Feldman.

Principles of Reverse Engineering

Regarding the task of reverse engineering of Boolean concepts Lee and Johnson-Laird (2013) postulates three principles related to difficulty. Those are the principle of variable components, the principle of positive outputs and the principle of dependence. Whereas the number of variable components is not applicable for this paper because all tasks of the experiment had the same amount of variables and thus can not be used to determine differences in difficulty, the other two principles appear to be more promising.

Principle of Dependence The principle of dependence takes the interdependency of the different variables into account. It states that "the greater the dependence of components on one another in determining the performance of the system, the harder the system should be to reverse engineer." (Lee & Johnson-Laird, 2013).

With respect to the lightswitch scenario this implies that if every single switch by its own is able to turn the light on and off, the respective components are considered independent. An example is a simple circuit with two switches connected as a disjunction as shown in Figure 3b. There, each switch can change the state of the light regardless of the state of the other switch. In contrast, in Figure 3a is an example for dependent components representing a simple conjunction combining two switches. There, each switch can only have an effect on the light if the other switch is in a certain state. Additionally, there is also the case of partial dependency, e.g., $a \wedge (b \vee c)$. In this case switch a is not able to turn the light on on its own because either switch b or switch c or both have to be on too, but a is capable of turning the light off independently from the state of the other switches.

Principle of Positive Outputs The third principle postulated by Lee and Johnson-Laird (2013), the principle of positive outputs, is based on the number of instances, i.e., different variable combinations that fulfil a given Boolean concept. The given example in Figure 4 has the Boolean concept of $(a \vee c) \wedge \neg\, b$ which is fulfilled by the three instances $a \neg\, b\, c$, $a \neg\, b \neg\, c$ and $\neg\, a \neg\, b\, c$ to turn on the light. Consequently the difficulty measure for this concept would be three.

Mental Models

The principle of positive outputs also is the foundation of the Mental Models approach (MM), which can be seen as an extension of the instances approach. It introduces a simplification of the instances to estimate the difficulty (Goodwin & Johnson-Laird, 2011). This idea is founded on the tendency of humans to eliminate unnecessary variables in their mental representations of Boolean concepts. To this end, the total number of instances is reduced by systematically eliminating irrelevant variables in order to merge two instances. The resulting simplified set of instances is considered to be an estimate of the mental models that participants have of the task.

Referring to the example in Figure 4, the three instances for the concept $(a \vee c) \wedge \neg b$ can be simplified (see Table 2). The only difference between the first two instances $a \neg b\, c$ and $a \neg b \neg c$ is the third variable c. Obviously if $a \neg b$ is given, the state of the third variable c is not important because it can be true or false but the light will still shine. Therefore, these two instances are simplified to $a \neg b$. However, the third instance can not be simplified any further, leading to a representation with two mental models. The difficulty is then estimated based on the number of mental models (e.g., 2 for the previous example).

Evaluation Data

The analysis of the present paper is based on the results of an experiment by Goodwin and Johnson-Laird (2011). For the research they used a modified experimental design which is based on the switch-task from Johnson-Laird (1983). The setup consists of three independent switches, similar to Figure 4, that control the light. They used nine concepts concerning

Table 2: Instances for an examplary Boolean concept and their corresponding Mental Models.

Boolean concept	Instances	Mental Models
$(a \vee c) \wedge \neg b$	$a \neg b\ c$	$a \neg b$
	$a \neg b \neg c$	$\neg a \neg b\ c$
	$\neg a \neg b\ c$	

three binary variables (switch *on* or *off*). These selected concepts were from a set of 250 possible concepts from Feldman (2003). Goodwin and Johnson-Laird (2011) chose the taken concepts observing their different complexities. In total, 28 students (12 male, 16 female) participated in the experiment. They were asked to describe the conditions in which the light turns on as a result of the positions of the three independent switches. At the beginning of every task, the switches were all turned off and the participants were presented with test trials to figure out which combinations turned the light on. To change the configuration of the switches they had to press a numbered button which was corresponding to the switch numbers. To see whether the light turned on or not the participants had to submit the configuration. Once they could describe the conditions in which the light turned on, they were able to press the "submit" button and proceed. They had to describe the conditions in their own words on a sheet of paper. During the experiment, participants were not allowed to take notes. If they were insecure about how to answer, they were instructed to describe as clearly as possible. Otherwise the response format was up to the participants.

The descriptions provided as responses by the participants considerably varied, but Goodwin and Johnson-Laird (2011) explained that assessing their accuracy (i.e., the correctness of the description) was straightforward. Two independent editors came to almost the same accuracies when interpreting the participants' descriptions.

While we are mostly relying on the original dataset, we augmented it by also annotating the direction of a description: Goodwin and Johnson-Laird (2011) found that, when describing the Boolean concepts, participants might switch from describing cases where the light would be turned on in the following to describing when the light would be turned off. In the following, we will refer to this as the *direction* of the description. Furthermore, the set of instances that cause the light to be turned on is referred to as the *onset*, while the *offset* denotes the set of instances causing the light to be turned off. According to Goodwin and Johnson-Laird (2011), participants might switch the direction in order to make the task easier, i.e., if the onset contains too many elements, a switch to the offset might occur. In the experiment by Goodwin and Johnson-Laird (2011) the change of direction was not explored any further. Still the given answers were considered correct when correctly relying on the *offset* instead of the *onset*.

Method

How good are the performances of the described models on an individual level? Compared to the experiment from Goodwin and Johnson-Laird (2011) the focus of the present paper was to find out how the previous presented models perform on an individual level. The following sections describe precise the analyses and results from the new analyses.

Goodwin and Johnson-Laird (2011) analyzed the previously described accounts for difficulty (Mental Models, Minimal Description Length and Algebraic Complexity) with respect to their ability to account for the difficulty of a concept. They assessed the capabilities of the approaches by comparing the correlations between the estimated difficulty of an approach with the average correctness achieved by participants. However, it remains unclear how the results would translate to an individual level, which will be investigated in the present paper. To this end, we implemented each of the presented approaches as an individualized model.

To facilitate this, we use the CCOBRA-framework[1] to ensure a modeling evaluation standard as proposed by Riesterer, Brand, and Ragni (2020b) with a focus on the models' capabilities to account for individual reasoning behavior. We relied on a coverage task, in which a model is presented with the complete set of information available for a specific individual reasoner, including the responses to all tasks (Riesterer, Brand, & Ragni, 2020a). This allows the model to fit to each reasoner, before it is then queried to replicate the responses for the tasks. To this end, it is important to note that this approach is not useful for testing data-driven models that can store the presented information, but, for cognitive models, provides insights into the model's ability to represent the reasoners response behavior in its parameter space. While it is an optimistic estimate of a models predictive capabilities, the correlation-based evaluations are also performed on the complete information. Therefore, we chose it as it can be seen as an extension of the correlation-based analysis to the individual level.

Each of our models consists of the core mechanism to estimate task difficulty (e.g., Mental Models) and a threshold that is used to decide at which point the difficulty is assumed to be too high for a specific participant (i.e., the difficulty at which the participant started to give incorrect answers). When fitted to an individual participant, the optimal value for the threshold was selected based on the accuracy to replicate the participant's correct responses and errors across all tasks.

Regarding the different approaches, we relied on fixed values for the tasks reported by Goodwin and Johnson-Laird (2011) for the *Minimal Description Length*, the *Algebraic Complexity* and the *Principle of Dependence* (referred to as *Dependency*).

The *Principle of Positive Outputs* was incorporated into a model (referred to as *Instances model*) that directly uses the number of instances as an estimate for the difficulty.

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

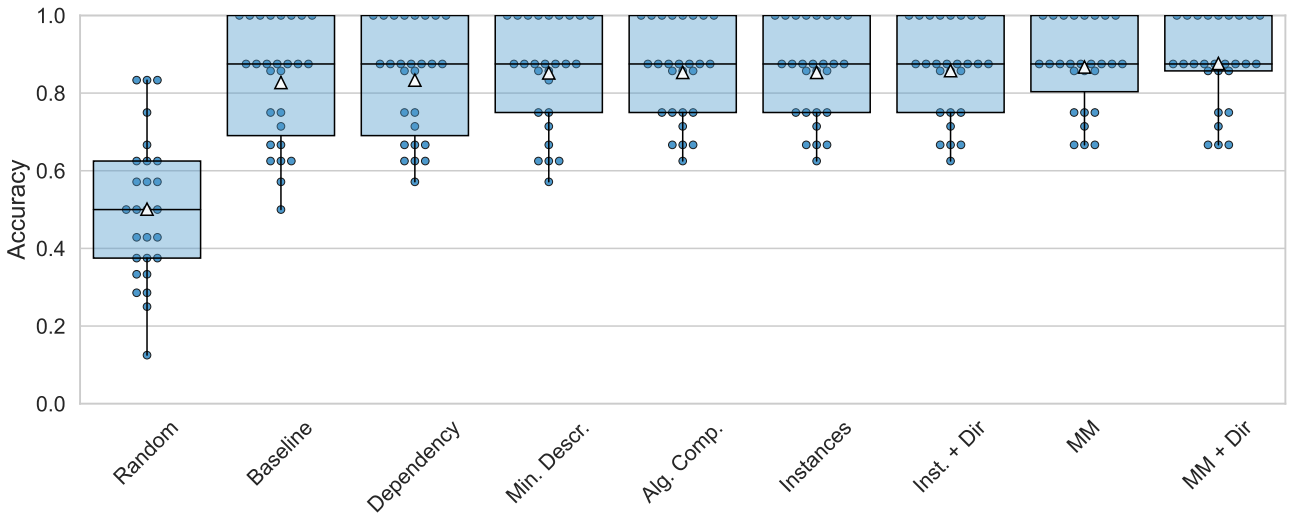


Figure 5: Detailed coverage performances of the models. The box-plots show the prediction accuracies of the implemented models with boxes ranging between the quartiles, the middle line indicating the median performance. The dots within each box-plot depict the single individuals. Triangles denote the mean accuracies.

Finally, the model based on the *Mental Models* approach (MM), relies on the *Instances model* to determine the initial set of instances. In order to rule out potential errors when implementing the simplification, our model then internally uses the original LISP model by Goodwin and Johnson-Laird (2011) for reducing the number of instances.

Two additional models were added as reference points for the performance: First the *Random* model, which determines the estimated correctness randomly (based on a uniform distribution) and can therefore be considered as a lower bound that any model should be able to surpass. Second, another *Baseline* model was included that assumes a perfect correctness by each participants. As the tasks are mostly solved correctly by the participants, it represents an aggregate model that does not consider individual differences. Therefore, it should be surpassed by any model that incorporates mechanisms to adapt to individuals.

Directions (Onset & Offset)

The previously introduced *direction* can serve as an extension for the *Instances model*, and therefore, also of the *Mental Models*. While MM focuses on the instances within the onset to determine the difficulty, the minimization process itself is agnostic of the direction. In a similar fashion, the *Instances model* could also rely on the number of instances in the offset instead of relying on the onset. In order to assess the effect of the direction, we used extended versions of the *Instances model* and the MM that rely on the onset or offset depending on the direction that the respective participant used for the given task. In the case that both directions were present in a participant's response, we used the onset as a default. The approach should be able to enhance the predictive capability of the Instances model and the MM by taking into account that the difficulty decreases in certain cases if the offset of instances is considered instead of the onset.

Results

Figure 5 shows the accuracy achieved by the models when replicating the participants' correctness. As expected, the *Random* model has the lowest performance with a mean and median accuracy of .5. While there are no differences in the median accuracy for all other models (*median* = .875), they differ with respect to their mean performance. All individualized models surpass the *Baseline* (*accuracy* = .825), which indicates that they can, at least to a small degree, reflect the individual correctness via the threshold. Overall, the approach based on *Mental Models* outperformed the other models, with a mean accuracy of .876, with the next best being the *Instances model* and the *Algebraic Complexity* with an accuracy of .85. While the *Minimal Description* model comes close (*accuracy* = .845), the *Dependency* model (*accuracy* = 0.83) barely surpasses the performance of the *Baseline* model. The additional information provided by including the direction could be used by both, the Instances model and MM. The improvement by the MM was higher (from *accuracy* = .866 to *accuracy* = .876) compared to the Instance model (from *accuracy* = .85 to *accuracy* = .855), but no substantial improvement was apparent. This, however, is more of a general problem: the peformance differences between the models were too small for any meaningful quantitative statement, as neither model was able to significantly outperform the baseline (Mann-Whitney-U between *MM + dir* and *Baseline*; $U = 306$, $p = .15$). This is likely due to the high ratio of correct responses and the low number of tasks available in the dataset, which leaves only very limited options for the models to set themselves apart from the others.

The amount of participants that achieved a perfect accuracy ($n = 8$), which could easily be replicated by all models, further reduced an already small dataset. To this end, even participants with only one mistake ($n = 9$) still do not allow for substantial differences in model performance. However, when considering the individual datapoints, it is possible to see that the individualization did in fact work. When comparing the lower quartile boundaries, it becomes apparent that the models show in fact differences for the individuals that did not always solve the tasks correctly.

Discussion

In the present article, several estimates for difficulty in Boolean concept tasks were evaluated. In contrast to Goodwin and Johnson-Laird (2011), our analysis was not performed on the basis of correlations between the estimate and the ground truth. Instead, we extended the different estimates to models that should account for the difficulty of a task with respect to an individual participant by introducing an additional threshold representing the maximum difficulty the participant could handle. We evaluated the models on the dataset from Goodwin and Johnson-Laird (2011). While the general trend found by our analysis was in line with the findings by Goodwin and Johnson-Laird (2011), the differences between the models were not significant. Especially when compared to a baseline model, that always assumes that participants solve a task correctly, a fundamental flaw of the dataset when used for model evaluation became apparent. A substantial amount of the participants (8 out of 28) solved every task correctly, with most other participants making only one or two mistakes. This meant that the models had only very limited possibilites to show any differences, which showed in the lack of any significant difference in terms of their performance.

However, some tendencies could be found nevertheless: When focusing on the lower quartiles, the models start to show differences, with the Mental Models having the edge. Furthermore, the inclusion of the direction, which was already expected to have an influence by Goodwin and Johnson-Laird (2011), did in fact allow the models to improve. MM was able to benefit more than the Instance model, which corroborates the assumption of MM that a simplification is in fact performed by reasoners.

From a more general perspective, the present analysis showed the importance of model evaluation on different settings, especially with a focus on individual participants. The different approaches differed substantially based on correlations alone, but did not translate to a more simulation-oriented setting, where a precise response to a task should match the response of a specific participant. To this end, the proposed evaluation with a well-defined setting and implemented, individualized models can serve as a first step.

Cognitive modeling should strive for the creation of models that are able to account for the human behavior, with as little interpretation and preprocessing of the recorded behavior as possible. The foundation to this also lies in a suitable data foundation, as model evaluation requires the ability to distinguish between different models. To this end, a suitable dataset should not only consist of a big corpus of participants, but should above all offer a large variety of tasks, which allows to find meaningful patterns in participants' responses. If the selected tasks are too easy or too difficult, evaluation will be impeded by ceiling/floor effects. In the setting of Boolean concepts, an extension of the tasks to tasks with more variables would also be important to add another dimension in which models and theories can differ. Furthermore, with a solid data foundation, the task can be extended from estimating the correctness into the task of predicting the precise description of the concept provided by a participant (in a standardized simplified way, e.g., by translating the description to a Boolean concept in a preprocessing step). Solving such a task, even in a simplified version, would require models to show a much deeper understanding of the reasoning processes that underlie solving the tasks.

References

Feldman, J. (2000). Minimization of boolean complexity in human concept learning. *Nature*, *407*(6804), 630–633.

Feldman, J. (2003). A catalog of boolean concepts. *Journal of Mathematical Psychology*, *47*(1), 75–89.

Feldman, J. (2006). An algebra of human concept learning. *Journal of mathematical psychology*, *50*(4), 339–368.

Goodwin, G. P., & Johnson-Laird, P. (2011). Mental models of boolean concepts. *Cognitive psychology*, *63*(1), 34–59.

Johnson-Laird, P. N. (1983). *Mental models: Towards a cognitive science of language, inference, and consciousness* (No. 6). Harvard University Press.

Lee, N. L., & Johnson-Laird, P. (2013). A theory of reverse engineering and its application to boolean systems. *Journal of Cognitive Psychology*, *25*(4), 365–389.

Love, B. C. (2002). Comparing supervised and unsupervised category learning. *Psychonomic bulletin & review*, *9*(4), 829–835.

Riesterer, N., Brand, D., & Ragni, M. (2020a). Do models capture individuals? Evaluating parameterized models for syllogistic reasoning. In S. Denison, M. Mack, Y. Xu, & B. C. Armstrong (Eds.), *Proceedings of the 42nd Annual Conference of the Cognitive Science Society* (pp. 3377–3383). Toronto, ON: Cognitive Science Society.

Riesterer, N., Brand, D., & Ragni, M. (2020b). Predictive modeling of individual human cognition: Upper bounds and a new perspective on performance. *Topics in Cognitive Science*, *12*(3), 960–974. doi: 10.1111/tops.12501

Shepard, R. N., Hovland, C. I., & Jenkins, H. M. (1961). Learning and memorization of classifications. *Psychological monographs: General and applied*, *75*(13), 1.

Smith, J. D., Minda, J. P., & Washburn, D. A. (2004). Category learning in rhesus monkeys: a study of the shepard, hovland, and jenkins (1961) tasks. *Journal of Experimental Psychology: General*, *133*(3), 398.