# How good can an individual's conclusion endorsement be predicted?

**Sara Todorovikj (todorovs@cs.uni-freiburg.de)**
**Marco Ragni (ragni@cs.uni-freiburg.de)**
Cognitive Computation Lab, Technical Faculty, University of Freiburg
Danish Institute for Advanced Study, University of Southern Denmark

## Abstract

Reasoning about conditional statements is relevant in science, culture, and our everyday life. It has been shown that humans do deviate from a classical logical interpretation of conditionals. Consequently, in the past years a number of cognitive models based on Bayesian or mental model approaches have been developed, whose performance is normally judged based on their ability to fit aggregate data of participants. Here, we diverge by focusing on the *individual* instead. Moreover, we propose a different model testing paradigm by analyzing on an existing large data set, how good current models are in *predicting* an endorsement of an individual reasoner on a scale from 0 to 100%. Towards this goal we reanalyze the data by rigorously distinguishing between test and training data set, by making existing models for conditional reasoning predictable such as the Dual Source Model (Singmann, Klauer, & Beller, 2016) and a model by Oaksford, Chater, and Larkin (2000). We also implement a modeling idea of Pearl based on possible worlds. We can show that all three models perform equally good in predicting an individual reasoner's endorsement and that they meet an empirical baseline (the median of the most frequent answer). A discussion on the gained insights in understanding conditional reasoning concludes the paper.

**Keywords:** Predictive modeling; cognitive modeling; conditional reasoning

## Introduction

In order to understand how human cognition works, a variety of cognitive models have been developed throughout the years and fitted to various experimental data. For example, consider the following reasoning task about a conditional statement (c.f., Singmann et al., 2016):

> If a balloon is pricked with a needle, then it will pop.
> A balloon is pricked with a needle.
>
> ---
>
> How likely is it that it will pop?

Your task would be to provide an answer between 0 and 100%. Now, imagine that a cognitive model is provided with the same task and makes a *prediction* of your response. Given experimental data, we propose that cognitive models are applied in such a predictive setting to each *individual*, as illustrated in Fig. 1. Comparing the true response and the prediction for all participants leads to a novel approach of cognitive model performance evaluation.

Motivated by the idea of Feynman that in order to fully understand something, one needs to be able to re-create it, Riesterer, Brand, and Ragni (2020) introduce a predictive modeling task in the syllogistic reasoning domain. They evaluated the predictive performance of syllogistic theories using

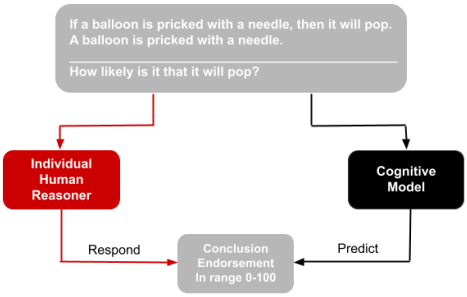


Figure 1: Predictive modeling task for endorsement rates

a modeling framework[1] called CCOBRA. Riesterer et al.'s (2020) focus is on the syllogistic domain, where a reasoner has only 9 answer options, meaning that a model either predicts the correct answer out of the possible 9 or not. This is where the scenario differs in our case. Here, we are dealing with a more complicated predictive task – one for *endorsements* that can be *any* value in the range 0-100. To understand the complexity of this task, consider the previously given reasoning task example once again. Since you are provided with the rule that if a balloon is pricked, it will pop, you would most likely gravitate towards answering with a 100%. But what happens, if:

> The balloon is without air, i.e., empty.

Then the balloon would *not* pop and you might give an answer that is less than 100%. Such aspects are called *disablers*. If an individual is aware of many disablers, their conclusion *endorsement* might be lower. On the other hand, there can be additional cases, called *alternatives*:

> A balloon can also pop, when it is pricked with something else than a needle.

Hence, depending on the cases different reasoners *have* in their minds, the given responses might differ. This introduces challenges when trying to predict how much a specific reasoner endorses a possible conclusion.

### Existing models and their comparison

In this paper we focus on the *conditional* reasoning domain. Conditionals are statements of the form "If X then Y" (also written as X→Y, where X is called the *antecedent* and Y the

[1] https://orca.informatik.uni-freiburg.de/ccobra/

*consequent*), often used to describe a causal relationship between two propositions X and Y. Given a conditional (called *major premise*) and a current state of a proposition (called *minor premise*), a *conclusion* can be inferred about the state of the other proposition. There are four inference forms: *modus ponens* (MP), *modus tollens* (MT), *affirming the consequent* (AC) and *denying the antecedent* (DA), as shown in Table 1.

Table 1: Inference Forms

| Name | MP | AC | DA | MT |
|---|---|---|---|---|
| Premise 1 | X→Y | X→Y | X→Y | X→Y |
| Premise 2 | X | Y | ¬X | ¬Y |
| Conclusion | Y | X | ¬Y | ¬X |

Singmann et al. (2016) studied the endorsements of the respective conclusions for the four inference forms in four experiments. Three of them focus on contents with varying amounts of disablers and alternatives. The fourth experiment introduces speaker expertise. We want uniform data, so we do not consider it. The authors also present a performance comparison of *Bayesian* modeling approaches for conditional reasoning. They are built upon the idea that a conditional "If X then Y" is understood by a conditional probability P(Y|X).

Oaksford et al. (2000) proposed that reasoning about a conditional rule can be modeled by the three parameters P(X), P(Y) and P(¬Y|X), the last one allowing for exceptions. Two extended versions using one and two additional exception parameters (Oaksford & Chater, 2007) and a model based on the Kullback-Leibler-Distance, have been statistically compared to a newly developed model – the Dual-Source-Model (DSM) – that assumes that individuals integrate two different kinds of processes: A knowledge-based component where they take Oaksford et al.'s (2000) approach and extend it with an additional form-based component, integrating both with a weight λ. We will present and explain the technicalities of these models in a following section.

While explicitly stating that model comparison should take model fit and model flexibility into account, due to the lack of Maximum Likelihood Estimation abilities for AIC/BIC, only a model fit using $R^2$ has been computed (Singmann et al., 2016). The $R^2$ goodness-of-fit values for the four models were used in a Linear Mixed Model (LMM) with random effects. Overall the DSM had the highest $R^2$, meaning that it was able to account for the highest percentage of variance, i.e., it had the best performance. So far the models have been evaluated on a statistical level given the analysis approach based on the $R^2$ and the LMM. In the following, however, we will focus on process aspects and – we will analyze if a model queried for a yet untested person is even able to predict an endorsement of a conclusion from 0 to a 100%.

## Our goal: Evaluating the *predictive* power of models

The current state of analysis does not convey yet, *if* the described models are predictive. When provided with observations on other participants' endorsement answers to a set of reasoning problems (= *training data set*), a cognitive model is called *predictive* for a (untested) reasoner, if it can correctly predict the inference endorsements (between 0 and 100%) for those problems (= *test data set*). This is rather easy for a yes/no question, as we only have two answer options for the model's prediction. However, it is *much more* challenging to develop a predictive setting for endorsement rates that range in the interval 0 - 100. Hence, this paper's first research question is: How can we *develop a predictive task setting and evaluate the predictions* and how can we adapt and evaluate the existing models to provide this prediction?

Our second research question – as current models are probabilistic – is it possible to have a cognitive model based on mental models? This is often questioned, as endorsement problems are usually considered *new paradigm*. Pearl has suggested approaches that combine a model structure with probabilities, which we will implement and compare too.

The paper is structured as follows: First, we present existing experimental data and Bayesian cognitive models for conditional reasoning. Second, we present an idea of Pearl, which we adapt to represent inference form endorsements. Third, we elaborate on how the benchmark was implemented. To conclude the paper, we present its predictive results, followed by a discussion.

## Data and Cognitive Models for Conditionals

We consider the experimental data provided in Singmann et al. (2016)[2], specifically the Experiments 1, 3a and 3b with 199 participants. In Exp. 3a and 3b, participants are divided in three groups. In two groups, participants are provided additional information in the form of alternatives and disablers, whereas the participants in the last group are provided only with the conditional task. All three experiments use the same four contents that have a varying amount of disablers and alternatives, both quantified with 'Few' and 'Many', shown in Table 2. The participants' task was to provide endorsement rates for the four inference forms as a probability in the range 0 - 100%. Each content is presented as a full conditional inference and as a reduced inference, i.e., no major premise, e.g., MP:

A balloon is pricked with a needle.

How likely is it that it will pop?

### Bayesian Cognitive Models

In the 60s a *deductive* path of cognitive modeling was followed, based on the assumption that logic is the basis for reasoning (Evans & Over, 2004). However, with time it has been shown that humans deviate from logic when given deductive reasoning tasks, and therefore, their responses are deemed false. That motivated the development of a new, Bayesian paradigm, where the models are based on probabilities and allow for background knowledge to be integrated when reasoning (Oaksford & Chater, 2020).

[2]The data can be found at `https://osf.io/zcdfq`.

Table 2: Contents used in Singmann et al. (2016) experiments.

| Keyword | Content | Disablers | Alternatives |
|---|---|---|---|
| Predator | If a predator is hungry, then it will search for prey. | Few | Few |
| Balloon | If a balloon is pricked with a needle, then it will pop. | Few | Many |
| Girl | If a girl has sexual intercourse, then she will be pregnant. | Many | Few |
| Coke | If a person drinks a lot of coke, then the person will gain weight. | Many | Many |

**Oaksford et al.'s (2000) model (OC)** Oaksford et al. (2000); Oaksford and Chater (2020) propose a probabilistic model for conditional reasoning. By using a $2 \times 2$ contingency table, as in Table 3, they represent conditional rules, where $a$ = P(X) and $b$ = P(Y), probabilities of the antecedent and consequent, respectively and $\varepsilon$ = P($\neg$Y|X) is the exception parameter.

Table 3: Contingency table for a conditional rule "If X then Y" Oaksford et al. (2000). There are three parameters: the probability of the antecedent P(X) denoted by $a$; the probability of the consequent P(Y) denoted by $b$; and a third parameter $\varepsilon$ for the probability of the exception P($\neg$Y|X).

| | Y | $\neg$Y |
|---|---|---|
| X | $a(1-\varepsilon)$ | $a\varepsilon$ |
| $\neg$X | $b-a(1-\varepsilon)$ | $(1-b)-a\varepsilon$ |

Derived from Table 3, this model uses the following equations for inference endorsement:

**MP:** P(Y|X) = 1 - $\varepsilon$ **DA:** $P(\neg Y|\neg X) = \frac{1-b-a\cdot\varepsilon}{1-a}$

**AC:** $P(X|Y) = \frac{a(1-\varepsilon)}{b}$ **MT:** $P(\neg X|\neg Y) = \frac{1-b-a\cdot\varepsilon}{1-b}$

As already mentioned, Oaksford and Chater (2007) present a more sophisticated version of this model. We decide to still take the original 2000 variant into consideration as the DSM builds up on it, as explained in the following.

**Dual-Source Model (DSM)** The DSM (Singmann et al., 2016) is an extension of Oaksford et al.'s (2000) model. It assumes that individuals integrate two different kinds of information: background knowledge about the content and information related to the logical form of the inference. It uses three types of parameters:

$\xi(C,x)$ – knowledge-based component, depending on the content $C$ and inference $x$, i.e. how much does an individual endorse an inference solely based on their background knowledge about the content

$\tau(x)$ – form-based component, reflecting the subjective probability of the inference form $x$, i.e. how much does an individual believe in the validity of an inference regardless of the content

$\lambda$ – a weight given to the form-based component (integrating $\xi(C,x)$ and $\tau(x)$ using Bayesian model averaging)

Applying the DSM to experimental data requires that participants have given endorsements to both a *reduced inference* and a *full conditional inference*. The model expresses the reduced inference endorsement through its knowledge-based component for content $C$ and inference $x$:

$$E_r(C,x) = \xi(C,x)$$

The $\xi(C,x)$ parameters are obtained by using Oaksford et al.'s (2000) equations, as shown above. Then, the endorsement of the full inference $x$ with content $C$ is given by:

$$E_f(C,x) = \lambda \cdot \{\tau(x) + (1-\tau(x)) \cdot \xi(C,x)\} + (1-\lambda) \cdot \xi(C,x)$$

The $\lambda$ parameter determines how much do individuals rely on form validity versus their background knowledge. $\tau(x)$ is the degree of belief in the full inference form. In case of uncertainties concerning the inference, the individual falls back to their background knowledge, through the weight (1 - $\tau(x)$) given to the knowledge-based component.

## Models and Probabilities: Applying an Idea of Pearl

**$\varepsilon$-semantics** Pearl (1991) introduced $\varepsilon$-semantics, a 'formal framework for belief revision', where belief statements are interpreted as statements of high probability and belief revision shapes current beliefs on newly available evidence. This approach seems to be most fruitful in our case, because disablers or alternatives can be such 'updates'. The idea of Pearl is based on the idea of possible worlds (or models) that can be assigned a probabilistic assignment (Pearl, 1991, p. 5):

> "Let $L$ be the language of propositional formulas, and let a *truth-valuation* for $L$ be a function $t$, that maps the sentences in $L$ to the set $\{1, 0\}$, (1 for 'true', 0 for 'false'). To define a probability assignment over the sentences in $L$, we regard each truth valuation $t$ as a world $w$ and define P($w$) such that $\sum_w$ P($w$) = 1. This assigns a probability measure to each sentence $l$ of $L$."

Before diving into our application of Pearl's idea, we will briefly touch upon mental models. A mental model consists of the truth states of the premise's propositions. Given a conditional premise "If X then Y", the initial mental model that an individual would construct is the one where both propositions are true, i.e. XY.

The Mental Model Theory (MMT) (Johnson-Laird & Byrne, 1991, 2002; Johnson-Laird, Khemlani, & Goodwin, 2015) assumes that once the initial model is created it triggers the recollection of relevant facts and knowledge. Those facts can either serve as evidence that the initial model is correct or will stimulate a search for alternatives leading to a second process where an extended mental model representation is obtained, also called a fleshed-out representation. It

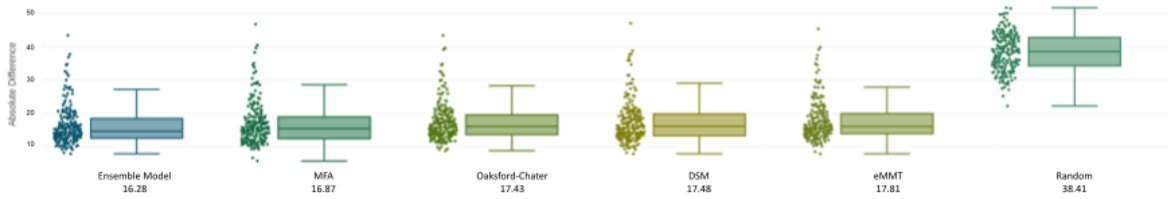


Figure 2: Boxplot depicting individual performance through the absolute difference between the predicted value and the true response. Overall mean absolute difference beneath the model's name. The comparable performance between OC, DSM and ε-MMT points to (a partial) functional equivalence.

contains models where X is false (¬X), as shown in Table 4a. This representation consists of all possible combinations of truth-values for X and Y for which the conditional "If X then Y" is true, which Johnson-Laird and Byrne (2002) call the *principle of truth*. This coincides with the material implication definition which is the leading interpretation of conditionals in the deductive paradigm.

## The ε-MMT

ε-MMT takes the mental model representation of *all* the conditional's propositions' truth state combinations, which we will refer to as *possible worlds*. In contrast to MMT, it also allows for the world X¬Y to exist, thus abandoning the material implication interpretation. Given a premise containing two propositions, X and Y, all possible worlds described by the premise along with the corresponding probability values are shown in Table 4b. Given a conditional "If a balloon is pricked with a needle then it pops", the probability of the world $\omega_2$, where the balloon is not pricked with a needle (X = 0) and it pops (Y = 1) is $p_2$.

Table 4: Representations of a conditional premise "If X then Y" with mental models and as possible worlds.

| Mental M. | Fleshed-out M. |
|---|---|
| X Y | X Y |
| ... | ¬X ¬Y |
| | ¬X Y |

(a) Johnson-Laird and Byrne (2002)

| World | X | Y | $P$ |
|---|---|---|---|
| $\omega_1$ | 0 | 0 | $p_1$ |
| $\omega_2$ | 0 | 1 | $p_2$ |
| $\omega_3$ | 1 | 0 | $p_3$ |
| $\omega_4$ | 1 | 1 | $p_4$ |

(b) Possible worlds, probability distribution $P$ and values $p_i$, $i \in (1,2,3,4)$

ε-MMT follows the same approach of previous accounts in the Bayesian paradigm, e.g. Oaksford et al. (2000), assuming that an individual's inference form endorsement can be expressed as a conditional probability of the conclusion given the minor premise.

$$P(\beta|\alpha) = \frac{P(\alpha \wedge \beta)}{P(\alpha)} \tag{1}$$

Following the definition of conditional probability, as shown in Eq.1 the four expressions shown below are obtained. They describe the endorsement of the four inference forms through the probability distribution $P$ of the conditional's worlds (Table 4b):

$$\textbf{MP:}\quad P(Y|X) = \frac{p_4}{p_3 + p_4} \qquad \textbf{DA:}\quad P(\neg Y|\neg X) = \frac{p_1}{p_1 + p_2}$$

$$\textbf{AC:}\quad P(X|Y) = \frac{p_4}{p_2 + p_4} \qquad \textbf{MT:}\quad P(\neg X|\neg Y) = \frac{p_1}{p_1 + p_3}$$

The parameters are bound by their sum, $\sum_i p_i = 1$, meaning that the number of free parameters for modeling one task is three. Total number of parameters to model an *individual* hence depends on the number of tasks they have to complete.

## Benchmark

In order to evaluate the three presented cognitive models, we implemented a benchmark within the framework CCOBRA, following Riesterer et al.'s (2020) approach. As already mentioned, their focus is on the syllogistic domain, where a model either predicts the correct answer out of the possible 9 or not, whereas our goal is to predict a value in the range 0-100. This poses a difficulty in adopting the same idea of judging a model based on whether it predicted the exactly correct answer or not. Instead, we are interested in *how close* the model's prediction is to the true reasoner's answer. The framework was extended to calculate the *absolute difference* between answers and predictions, rather than check for accuracy. In their benchmark, a theory is considered to have a good performance if it has a high accuracy rate. In our case, a cognitive model aims for a low absolute difference.

Generally, in order for the cognitive model to be able to predict a response as accurately as possible, it needs to be exposed to already existing data, i.e. a training set, from which it can learn. Here, we used Singmann et al.'s (2016) experimental data which we presented earlier. Since all three experiments have the same contents, we combined their data into one set, as the authors did in their original study. When provided with the same data for both training and testing, as in our case, the CCOBRA framework uses a leave-one-out cross-validation method – models are fitted on every participant, except the one whose answers are to be predicted. The same process is repeated for each participant.

In the training phase, we fit the models to the participants'

Table 5: Medians of the models' parameters per task and conditional presentation form. Values discussed below are in bold.

| Form | Task | ε-MMT[1] | | | | Oaksford-Chater[2] | | | Dual-Source Model[3] | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | $p_1$ | $p_2$ | $p_3$ | $p_4$ | $a$ | $b$ | ε | $\xi(C,\text{MP})$ | $\xi(C,\text{AC})$ | $\xi(C,\text{DA})$ | $\xi(C,\text{MT})$ |
| Reduced Inference | Predator | .44 | .06 | .03 | .59 | .56 | .60 | .10 | .90 | .85 | .80 | .86 |
| | Balloon | .48 | **.12** | .05 | .38 | **.38** | .48 | .12 | .88 | **.70** | **.77** | .91 |
| | Girl | .23 | .06 | **.42** | .13 | .63 | **.23** | **.68** | **.33** | .87 | .92 | **.45** |
| | Coke | .27 | **.20** | **.14** | .29 | .47 | .53 | **.37** | **.63** | **.56** | **.55** | **.63** |
| Conditional Inference | Predator | .46 | .05 | .02 | .62 | .56 | .59 | .08 | | | | |
| | Balloon | .49 | **.07** | .02 | .53 | **.46** | .53 | .08 | | | | |
| | Girl | .32 | .06 | **.22** | .35 | .60 | **.41** | **.39** | | | As above | |
| | Coke | .33 | **.15** | **.06** | .40 | .47 | .55 | **.23** | | | | |

[1] Probabilities of possible worlds $\omega_i$, see Table 4b, we have $p_1 = \text{P}(\omega_1)$, $p_2 = \text{P}(\omega_2)$, $p_3 = \text{P}(\omega_3)$, $p_4 = \text{P}(\omega_4)$, note that only 3 parameters are necessary because of $\sum_i p_i = 1$; [2] The three parameter values are: $a = \text{P(X)}$, $b = \text{P(Y)}$, $\varepsilon = \text{P}(\neg\text{Y}|\text{X})$; [3] Knowledge-based parameters $\xi(C,x)$ for content $C$ and inference $x$. The same values are used in the conditional case.

answers by optimizing the models' parameter values such that the absolute difference between the predicted answer and the reasoner's response is minimized. In order to do that we used Python's `scipy.optimize.minimize`[3] with the method Sequential Least Squares Programming (SLSQP). This method was chosen because it allows for constrained minimization.

Following Riesterer et al. (2020), we included a Random model as a lower bound, which in our case gives a random value in the range 0-100 as a prediction. Our models do not adapt to the individual, so we also included a Most Frequent Answer (MFA) model as an upper bound. In old paradigm experiments such a model would count the number of times an inference has been accepted or rejected and would predict the outcome that was most frequent. However, now we have a far more complex situation, dealing with a big range of values, to which we had to adapt this idea by having the MFA model give the *median* of the responses as a prediction.

## Predictive Modeling Results

We judge a model's performance by the mean of the absolute differences between the model's predictions and the individuals' answers. A lower absolute difference indicates more accurate predictions and therefore, better performance.

Figure 2 illustrates the model performance for each individual. The probabilistic models have similar results, all three greatly outperforming the Random model, while being comparable to the MFA model, but not better. OC and the DSM, both established models in the current Bayesian paradigm give an impressive performance. But, now we can also see that ε-MMT, a *model-based approach* is a valuable competitor in this probabilistic paradigm.

Having a predictive performance that is as good as an empirical model is an accomplishment for the probabilistic cognitive models. However, if we compare only the three of them – their performance is not very different. So, we proceed with the analysis by investigating the models' parameter values and how they aid in explaining the individuals' conditional interpretations. The median values of the models' parameters are shown in Table 5. In the reduced inference case participants are not provided with a rule, so their background knowledge is more prominent and that is reflected in the parameter values. We use now X for the antecedent of a conditional, and Y for its consequent ("If X then Y"). In the case of ε-MMT, the parameter $p_2$ describes the probability of the world $\omega_2$ where Y happens even if X does not and its values are higher for tasks with 'Many' alternatives, in contrast to 'Few'. The parameter $p_3$, on the other hand, is the probability of the world $\omega_3$ where X is true, however Y is not and through higher values shows the presence of 'Many' disablers. It can be seen how when a conditional has been provided, the belief in these two worlds diminishes. For OC, the most noticeable impact is on the ε parameter which is the probability of the exception $\text{P}(\neg\text{Y}|\text{X})$. Its values are exceptionally higher for tasks with 'Many' disablers. A lower value for $a = \text{P(X)}$ is present in the case of 'Many' alternatives, showing that X does not need to be true for Y to happen. Likewise, $b = \text{P(Y)}$ reflects the presence of 'Many' disablers which would prevent Y from occurring. The influence of alternatives and disablers is reflected in the conditional case as well, though at a smaller scale due to the conditional rule restricting the integration of background knowledge, similarly to ε-MMT. For the DSM, we have the four knowledge-based parameters $\xi(C,x)$ for each content $C$ and inference form $x$. Their values correspond to the inference form endorsements in the reduced inference case. Alternatives suppress the logically invalid forms, AC and DA, which is shown through $\xi$'s values for the tasks with 'Many' alternatives. Similarly, as disablers suppress the logically valid MP and MT, the corresponding $\xi$ values for tasks with 'Many' disablers are noticeably lower. The other parameters have the following median values: $\tau(\text{MP}) = 1.00$, $\tau(\text{AC}) = 0.40$, $\tau(\text{DA}) = 0.49$, $\tau(\text{MT}) = 0.88$ and $\lambda = 0.78$. Larger values of $\tau$ for MP and MT show higher beliefs in the logically valid forms MP, MT.

Considering each individual from 199 participants, 81 were best predicted by OC, another 81 by the DSM and 37 by ε-MMT. That lead us to the conclusion that among these three models, there is not a single one that "dominates" the others. Therefore, in order to support the idea that one single model can *not* capture *every* individual, we combined all

[3] `https://docs.scipy.org/doc/scipy/reference/generated/scipy.optimize.minimize.html`

three models into what we call an ensemble model, using for each individual the model that made the best prediction. This model consists of the best that these cognitive models can offer and it outperformed the MFA Model by reaching a mean deviation of ca. 16%. The purpose of this ensemble approach is to show that integrating strategies captures individuals best.

## Discussion and Conclusion

Our first research question was if it is possible to predict a conclusion endorsement varying in the range between 0 and 100, and not a dichotomous "yes" or "no" response. Yes – our results show that two Bayesian models exposed to a training data set can generate predictions on unseen data with a mean deviation of 17%. With this, we are establishing a new testing paradigm by not asking how good are the models in explaining existing data, but rather how good can they *predict* a reasoner's answers? By doing that we can elicit new insights. E.g., Singmann et al. (2016) showed that the DSM outperforms other probabilistic models when comparing their fits. However, evaluating the predictive power, the DSM does not perform better than the model it is built upon, the Oaksford et al.'s (2000) original probabilistic model, meaning that in this task only that one source is enough. We posed the question whether it would be possible that a model-based approach could compete with Bayesian models. Elqayam and Over (2013) discuss how old paradigm theories, like the MMT, focus on truth preservation from assumptions and cannot account for irrationality in human decision making. Here we took MMT's conditional representation and adapted it such that it does not follow the old paradigm's material implication interpretation and extended it with probabilities based on Pearl's (1991) ε-semantics. With that, we showed that – yes, a model-based approach can indeed compete with established Bayesian models. The comparable performance of the 3 cognitive models indicate a functional equivalence and similar processes, but, they do differ in their representation. None of the single models predictive performance was better than the MFA. This has been regarded in other domains such as syllogistic reasoning as an empirical upper bound for static models (Riesterer et al., 2020). By combining them into a ensemble model and introducing a better representation flexibility we showed that this performance upper bound can be surpassed while still having the tools to give insight into individuals' conditional reasoning, capturing individual differences. By looking into the models' parameter values we learn how disablers and alternatives influence the reasoners' representation of the conditional from different perspectives. Consider a task with 'Many' disablers, through ε-MMT's $p_3$ parameter we understand that the individual's belief in the world $\omega_3$, where the antecedent has happened but the consequent has not, is stronger. OC shows us that individuals assign a high probability to the conditional's exception, $P(\neg Y|X)$. The DSM shows through its $\xi$ parameters how disablers suppress the logically valid MP and MT, which is a reasoning effect that has been long recognized in this field (Byrne, 1989). We took into consideration experiments that deal with meaningful contents. Data is (still) quite scarce, as the focus in experiments has largely been on reasoning about abstract material. Our interest is in how humans reason in their *everyday life*, where most of our reasoning takes place. Hence we use such material. Nonetheless, the methods can be applied to abstract problems too.

This work opens future research lines in comparing how parts of models can be translated into each other. It not only allows to ground some of the *functional equivalence* we have already identified, but it would additionally help recognize where models deviate and what reasoning strategies might be missing when modeling an individual. With that, predictions of the reasoner's conclusion endorsement would improve, which would lead to a better understanding of the reasoning processes, making this path of not only fitting models, but also challenging their predictive capabilities an exciting one, opening many doors to a new way of adaptive modeling.

## Acknowledgements

This work was supported by DIAS and grants MR 1934/4-1 and MR 1934/10-1.

## References

Byrne, R. M. (1989). Suppressing valid inferences with conditionals. *Cognition*, *31*(1), 61.

Elqayam, S., & Over, D. E. (2013). New paradigm psychology of reasoning: An introduction to the special issue edited by Elqayam, Bonnefon, and Over. *Thinking & Reasoning*, *19*(3-4), 249.

Evans, J., & Over, D. E. (2004). *If*. Oxford University Press.

Johnson-Laird, P. N., & Byrne, R. M. (1991). *Deduction*. Lawrence Erlbaum Associates, Inc.

Johnson-Laird, P. N., & Byrne, R. M. (2002). Conditionals: a theory of meaning, pragmatics and inference. *Psychological review*, *109*(4), 646.

Johnson-Laird, P. N., Khemlani, S. S., & Goodwin, G. P. (2015). Logic, probability and human reasoning. *Trends in cognitive sciences*, *19*(4), 201.

Oaksford, M., & Chater, N. (2007). *Bayesian rationality: The probabilistic approach to human reasoning*. Oxford University Press.

Oaksford, M., & Chater, N. (2020). New paradigms in the psychology of reasoning. *Annual Review of Psychology*, *71*, 305.

Oaksford, M., Chater, N., & Larkin, J. (2000). Probabilities and polarity biases in conditional inference. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *26*(4), 883.

Pearl, J. (1991). *Epsilon-semantics* (Tech. Rep.). Computer Science Department, University of California.

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

Singmann, H., Klauer, K. C., & Beller, S. (2016). Probabilistic conditional reasoning: Disentangling form and content with the dual-source model. *Cognitive Psychology*, *88*, 61.