Putting the Probability Heuristics Model to the Test

Lukas Elflein (elfleinl@cs.uni-freiburg.de) and **Marco Ragni (ragni@cs.uni-freiburg.de)**
Cognitive Computation Lab, Department of Computer Science, Albert-Ludwigs-Universität Freiburg
Georges-Köhler-Allee 52, 79110 Freiburg, Germany

Abstract

In the last decades there was a shift from more logically inspired theories describing human reasoning towards the new paradigm of probabilistic approaches. One of the most prominent models for syllogistic reasoning is the Probability Heuristics Model (PHM) which has been formulated based on five heuristics. The contribution of this article is: (i) to provide an analysis of different formalizations of the PHM, (ii) to examine the impact of each heuristic, and (iii) to identify possible violations of underlying assumptions in present implementations. A systematic analysis of the model parameters shows a surprising variation in parameter values across experiments. A Bayesian modeling approach explains this variance of parameters. Implications for probabilistic approaches are discussed.

Keywords: Syllogistic Reasoning; New Paradigm of Reasoning; Cognitive Modeling; Heuristics; Bayesian; PHM

Introduction

A classical *syllogism* consists of two quantified statements (sometimes called premises) using one of the quantifiers All (abbreviated by A), Some (I), Some ... not (O), or None (E). And the task is to draw an inference from the given information, if possible. Consider the following example:

Premise 1:	All a are b (Aab)
Premise 2:	Some b are c (Ibc)

What, if anything, follows?

Most participants [74%, (Khemlani & Johnson-Laird, 2012)] conclude that *Some a are c* (Iac) follows. The given conclusion is, when evaluated with first-order logic, not correct. If we allow only answers between the two terms 'a' and 'c' (in either direction), the 4 quantifiers above, or *no valid conclusion* (NVC), then only the latter is correct in terms of first-order logic. Four different arrangements of the terms in the premises are possible. These are called figures, where we use the numbering of the figures in (Khemlani & Johnson-Laird, 2012) notation:

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

And so any syllogism can be described by the respective quantifiers of the first and second premise and the Figure. For the example above we can simply write: AI1 to uniquely characterize the syllogism. While the human responses in experiments often deviate from the inferences of first-order logical calculus, Chater and Oaksford (1999) propagated a *new paradigm of reasoning* that is inspired by a Bayesian interpretation of cognition. A core idea is that different quantifiers have a different degree of informativeness I, i.e., conclusions containing specific quantifiers are more informative than others. The ordering is given by:

$$I(\mathsf{A}) > I(\mathsf{I}) > I(\mathsf{E}) > I(\mathsf{O}) \tag{1}$$

The PHM has been formulated based on this ordering (Chater & Oaksford, 2007) by three *generation heuristics* (G for short) predicting the responses of a human reasoner (more are possible due to reasoning errors):

G1 The min-heuristic: Choose the quantifier of the conclusion to be the same as the quantifier in the least informative premise (the *min-premise*). In the example above (AI1) the min-premise is *Some b are c*, and so the min-quantifier in the conclusion is *Some* (I).

G2 p-entailments: The next most preferred conclusion will be the p-entailment [see eq. (8)] of the quantifier predicted by the min-heuristic. In the example above (AI1), the min-quantifier is *Some* (I), which entails *Some ... not* (O).

G3 Attachment-heuristic: If just one of the possible conclusion subject noun phrases matches the subject noun phrase of just one premise, then the conclusion has that subject noun phrase. In the example above (AI1) this heuristic predicts the terms in the conclusion as a-c.

Furthermore, there are two *test heuristics* motivating a parametrization of the possible responses:

T1 Max-heuristic: Be confident in the conclusion generated by G1-G3 in proportion to the informativeness of the most informative premise (the max-premise).

T2 O-heuristic: Avoid producing or accepting O-conclusions, because they are uninformative relative to other conclusions.

From these heuristics we can observe that: (i) A reasoner responding 'No valid conclusion' is not predicted by PHM. Hence, the question arises how to deal with such responses and if an extension of PHM is necessary (like one (Hattori, 2016) proposed). (ii) Different heuristics can generate conflicting predictions. Consider as an example the following syllogism (AO1): All a are b and Some b are not c. Then the min-heuristic (G1) predicts the O-conclusion *Some a are not c* responses, and the prediction is confident, because of the A-quantifier and the max-heuristic (T1), but the O-heuristic (T2) calls for avoiding the O-quantifier. Testing the coherence of the PHM becomes important. (iii) A systematic analysis of parameters and their stability to evaluate the goodness-of-fit on diverse data set including individual data. (iv) A rigorous

mathematical model allows to evaluate PHM and to develop a Bayesian implementation of the theory.

The heuristics (of the PHM) have been implemented by Chater and Oaksford (2007) with 6 parameters: one for each quantifier ($\{p_{\mathsf{A}}, p_{\mathsf{I}}, p_{\mathsf{E}}, p_{\mathsf{O}}\}$), an entailment probability (p_{ent}), and an error probability (p_{err}). To fit these parameters to experimental data, different strategies can be employed. Originally, Chater and Oaksford "obtained best fit estimates of these six parameters directly from the mean values of the relevant quantities in the data". This procedure minimizes some deviation on the 256 datapoints excluding the NVC responses. Recently, it has been implemented by (Hattori, 2016) with an additional fit for NVC and tested on 8 studies (aufzaehlen...) for predicting the quantifiers or the NVC response. Hence there are two implementations of PHM and to compare the data on common grounds we use the 4 studies not reporting 'Misc' answers or 'Most'/'Few' quantifiers. We abbreviate these by the individual names of the studies, BBJ95 (Bara, Bucciarelli, & Johnson-Laird, 1995), JB84-3 (Johnson-Laird & Bara, 1984), JS78-2a (Johnson-Laird & Steedman, 1978), JS78-2b (Johnson-Laird & Steedman, 1978). Additionally, to examine figurative effects and to fit individual data we use a data set that has been published with the 2017 syllogistic modeling challenge[1]. It consists of 139 participants that solved all 64 syllogisms in a paid web-experiment. Participants received the premises with the same content used in (Chater & Oaksford, 1999) and had to select the respective response. We abbreviate the data (that we will use to calculate individual performance) by RG16. The remainder of the article is structured as follows: In the next section we propose a formalization of the PHM by mathematical equations. In Section 3 we demonstrate the equivalence of the previous PHM implementations and our formalization by reproducing the reported results from (Chater & Oaksford, 1999) and (Hattori, 2016). This includes discussing the different parameters. In Section 4 we show that using a full Bayesian account allows to answer new questions. Limitations and potentials of our formal approach including implications for PHM are discussed.

A New Mathematical Formalization of PHM

The goal of this section is to develop a mathematical framework of the PHM.

PHM as a Binary Decision Diagram. We are first representing the PHM as a binary decision diagram (see Fig. 1). It helps to make the cognitive processes transparent and to introduce some necessary notation. The attachment heuristic does not depend on the G1 or G2 heuristic. Hence, if the answer is compatible with the attachment heuristic (G3), proceed to the min-heuristic (G1). If the quantifier of the conclusion is the min-quantifier, accept it with probability $p_{Q_{max}}$. If the min-heuristic fails, check if the entailment of the min-quantifier matches (G2). Accept this conclusion with p_{ent} if the entailment heuristic is satisfied, else use p_{err}. For example, for the AI1-syllogism, the answer Oac satisfies G3, fails G1, but satisfies G2, thus it get assigned a probability of p_{ent}. Ica satisfies G1 and G2, thus the probability is $p_{Q_{max}}$. The max-quantifier is A, thus $p_{Q_{max}} = p_A$. All other answers are accepted with probability p_{err}.

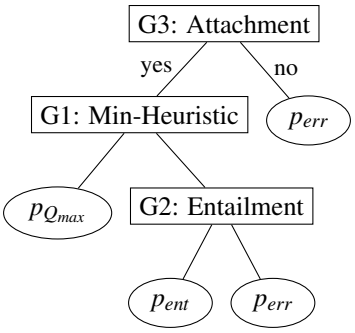

Figure 1: PHM decision flowchart for forced choice syllogisms with 9 possible answers. For every possible answer, its compatibility with the generating heuristics is checked and assigned a probability $p_{Q_{max}}$, p_{ent}, or p_{err}.

Closed-form equation. In this subsection we demonstrate how the PHM can be formulated by a closed-form equation. The information about a syllogism can be written as a set of two premises $\{(D_0, Q_0), (D_1, Q_1)\}$ with a direction (also called figure) D_i and quantifier Q_i of each syllogism, the ordering of the premises is not considered. The quantifiers are again A, I, E, O. To represent the direction we introduce a predicate D encoding the order of the terms (for simplicity we assume that the middle term in the premises is always 'b'), so

$$D = \begin{cases} 1, & \text{for ab, bc, ac} \\ 0, & \text{for ba, cb, ca} \end{cases} \tag{2}$$

e.g., the AI1 syllogism 'All a are b, some b are c' can be encoded by $\{(1,A),\ (0,I)\}$. The conclusion consists of a direction D_c and a quantifier Q_c. **The min-quantifier** Q_{min} is the quantifier among the premises with the least informativeness (cp. heuristic G1 above). So with the informativeness-ordering taken from Chater and Oaksford (1999) (see equation (1) above) we set

$$Q_{min} = \underset{i}{\operatorname{argmin}}(I(Q_i)) \tag{3}$$

The index i is used to make Q_i reference the respective quantifier. In the following we will use the **delta function**. In is defined as 1 if its indices are equal, otherwise it is 0:

$$\delta_{x,y} = \begin{cases} 1, & \text{for } x = y \\ 0, & \text{else} \end{cases} \tag{4}$$

Now we can use this delta-function to mimic the **min-heuristic**, it evaluates to 1 if the conclusion quantifier equals

[1] http://www.cc.uni-freiburg.de/data/rg16

the min-quantifier, otherwise to 0:

$$\delta_{Q_c,Q_{min}} = \begin{cases} 1, & \text{for } Q_c = Q_{min} \\ 0, & \text{else} \end{cases} \quad (5)$$

This min-heuristic can be combined with the **max-heuristic** to obtain the probability of a participant answering with some quantifier Q_c. This conditional probability depends on the human reasoner receiving the premise information and possible conclusions (symbolized with $|Data$).

$$P(Q_c|Data) = \delta_{Q_c,Q_{min}} \cdot p_{Q_{max}} \quad (6)$$

$$\text{with } Q_{max} = \text{argmax}_i(I(Q_i)) \quad (7)$$

The probabilities $p_{Q_{max}}$ actually are 4 fitting parameters, one for each quantifier ($p_\mathsf{A}, p_\mathsf{I}, p_\mathsf{E}, p_\mathsf{O}$). In the original PHM framework (Chater & Oaksford, 2007), there is the possibility of answering with the entailment of the min-quantifier $Ent(Q_{min})$, the **entailment heuristic (G2)**. Originally, there is only one parameter describing its frequency, p_{ent}. Thus, the entailment heuristic is independent of the max-premise, and therefore does not follow the max-heuristic, which would require 4 separate parameters or some other dependency on the max-premise. Thus the original implementation is incompatible with its verbal description (described above). We will follow the original implementation instead of the original verbal description in the following. The entailment function (i.e., which quantifier is entailed by the current quantifier) is taken from Chater and Oaksford (1999).

$$Ent(\mathsf{A}) = \mathsf{I}, \quad Ent(\mathsf{I}) = \mathsf{O}, \quad Ent(\mathsf{O}) = \mathsf{I}, \quad Ent(\mathsf{E}) = \mathsf{O} \quad (8)$$

This entailment is modeled by probability p_{ent} computed on experimental data. By combining the min/max-heuristic with the entailment-heuristic, we can formulate an equation describing that a reasoner choses a quantifier either according to the min-max or to the entailment heuristic:

$$P(Q_c|Data) = \delta_{Q_c,Q_{min}} \cdot p_{Q_{min}} + \delta_{Ent(Q_{min}),Q_c} \cdot p_{ent} \quad (9)$$

Finally, we add the possibility of erroneously accepting another quantifier with an **error rate** p_{err}. Chater and Oaksford (1999) choose to let errors occur if no other heuristic is applied. In our modeling we use an inverse delta-function $(1-\delta_{x\,y})$ to select the cases where neither the entailment nor the min-max heuristic is used:

$$P(Q_c|Data) = \delta_{Q_c,Q_{min}} \cdot p_{Q_{max}} + \delta_{Ent(Q_{min}),Q_c} \cdot p_{ent} + (1-\delta_{Q_c,Q_{min}}) \cdot (1-\delta_{Ent(Q_c),Q_{min}}) \cdot p_{err} \quad (10)$$

To include effects of the directions, we need to formulate a mathematical equation for the **attachment heuristic (G3)**. There are at least two different descriptions (Chater & Oaksford, 1999, 2007), we follow the more recent one. It is not clear how the Figures 3 and 4 (see the Introduction) need to be treated. Accepting both directions in the conclusion with equal probability seems like a plausible interpretation. On the other hand, the direction of the conclusion D_c for figure 1 and 2 is clear:

$$P(D_c|Data) = \delta_{D_0,D_1}\delta_{D_c,D_1} + (1-\delta_{D_0,D_1}) \cdot 0.5 \quad (11)$$

If the conclusion direction is not predicted by attachment, we still accept it with error rate p_{err}:

$$P_{D\ error} = \delta_{D_0,D_1}(1-\delta_{D_c,D_1}) \cdot p_{err} \quad (12)$$

By combining equations (10), (11), and (12), we get the following formula for PHM:

$$P(C|Data) = P(D_c|Data) \cdot P(Q_c|Data) + P_{D\ error} \quad (13)$$

$$\begin{aligned} &= [\delta_{D_0\ D_1}\delta_{D_c\ D_1} + (1-\delta_{D_0\ D_1}) \cdot 0.5] \\ &\quad \cdot \left[\delta_{Q_c\ Q_{min}} \cdot p_{Q_{max}} + \delta_{Ent(Q_{min})\ Q_c} \cdot p_{ent} \right. \\ &\quad \left. + (1-\delta_{Q_c\ Q_{min}}) \cdot (1-\delta_{Ent(Q_c)\ Q_{min}}) \cdot p_{err}\right] \\ &\quad + (1-\max_i(\{\delta_{D_c\ D_i}\}_i)) \cdot p_{err} \end{aligned} \quad (14)$$

Evaluating the Formal Model We need to show that our formalization captures the results computed by Chater and Oaksford (1999). If we neglect the figures like they do, we can insert the 16 moods and 4 quantifiers into the formula. Thus we obtain a 64×4 matrix for the syllogistic answers, where 4 lines are equal due to not considering the figure. Setting the parameters to the values provided by Chater and Oaksford (1999) (denoted by $\stackrel{*}{=}$), the resulting matrix of numbers is identical to the 'Model' columns of their table:

$$\begin{bmatrix} p_A & p_{ent} & p_{err} & p_{err} \\ p_A & p_{ent} & p_{err} & p_{err} \\ \vdots & \vdots & \vdots & \vdots \\ p_{err} & p_a & p_{err} & p_{ent} \\ \vdots & \vdots & \vdots & \vdots \\ p_{err} & p_{ent} & p_{err} & p_O \end{bmatrix} \stackrel{*}{=} \begin{bmatrix} 70.14 & 10.76 & 1.22 & 1.22 \\ 70.14 & 10.76 & 1.22 & 1.22 \\ \vdots & \vdots & \vdots & \vdots \\ 1.22 & 70.14 & 1.22 & 10.76 \\ \vdots & \vdots & \vdots & \vdots \\ 1.22 & 10.76 & 1.22 & 18.04 \end{bmatrix}$$

The full PHM model given in equation (14) can be represented by an analogous 64×8 matrix. For each max-quantifier, a separate parameter for the 'NVC' answer can be added to normalize the rows to 1, which has technical advantages during fitting and model comparison:

$$p_{Q_i} + p_{ent} + 6 \cdot p_{err} + p_{nvc}^{Q_i} = 1$$

$$p_{nvc}^{Q_i} = 1 - p_{Q_i} + p_{ent} + 6 \cdot p_{err}$$

Giving us a 64x9 matrix:

$$\begin{bmatrix} p_a & p_{ent} & p_{err} & p_{err} & p_{err} & p_{err} & p_{err} & p_{err} & p_{nvc}^A \\ p_{err} & p_{err} & p_{err} & p_{err} & p_a & p_{ent} & p_{err} & p_{err} & p_{nvc}^A \\ \frac{p_a}{2} & \frac{p_{ent}}{2} & p_{err} & p_{err} & \frac{p_a}{2} & \frac{p_{ent}}{2} & p_{err} & p_{err} & p_{nvc}^A \\ \frac{p_a}{2} & \frac{p_{ent}}{2} & p_{err} & p_{err} & \frac{p_a}{2} & \frac{p_{ent}}{2} & p_{err} & p_{err} & p_{nvc}^A \\ p_{err} & p_a & p_{err} & p_{ent} & p_{err} & p_{err} & p_{err} & p_{err} & p_{nvc}^A \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \\ p_{err} & \frac{p_{ent}}{2} & p_{err} & \frac{p_o}{2} & p_{err} & \frac{p_{ent}}{2} & p_{err} & \frac{p_o}{2} & p_{nvc}^O \end{bmatrix}$$

Empirical testing

Improved frequentist fitting. Chater and Oaksford (1999) originally "computed the fit between data and model over all five responses, A, I, E, O, and NVC [...]. Estimates for the NVC response did not involve introducing any further parameters because they can be derived from a linear combination of the existing parameters." To do this, a standard method is optimizing the root mean square error on all aggregated syllogistic data.

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i^{mod} - y_i^{exp})^2} \tag{15}$$

$$RMSE_{Hattori} = \frac{1}{64}\sum_{j=1}^{64}\sqrt{\frac{1}{5}\sum_{i=1}^{5}(y_{ij}^{mod} - y_{ij}^{exp})^2} \tag{16}$$

Similarly, Hattori (2016) used a mean of the same metric (eq. 16). Optimizing the RMSE (eq. 15) via BFGS (python code is provided[2]) improves the fit compared to the fits of Oaksford and Chater (see Table 1). The best fit values are slightly different from those given by Chater and Oaksford (1999) (also compare Figure 3).

Table 1: Optimizing parameters with a root mean square error (*RMSE*) improves goodness of fit compared to the original method (CO99). CO99 is similar to optimizing a RMSE on all data except the NVC column ($RMSE_{-NVC}$).

Method	RMSE	p_a	p_i	p_e	p_o	p_{ent}	p_{err}
CO99	.101	.70	.31	.19	.18	.11	.01
$RMSE_{-NVC}$	.101	.70	.31	.19	.18	.11	.02
RMSE	.099	.68	.35	.18	.17	.11	.02

Contribution of heuristics

We want to quantify the contributions of every single heuristic independently. Equation 6 already defines the min-heuristic. Thus, we can compute the impact of the min-heuristic (G1) alone or in combination with the entailment-heuristic (G2) to determine how often the conclusion quantifier is predicted. Additionally, we can test the contribution of the attachment heuristic (G3) and compare it to uniform guessing of the direction of the conclusion. To do this, we need information on the direction of the conclusion. This information is not present in the Hattori (2016) data, but it is in the RG16 data. Thus, an aggregation of the RG16 data is used for comparing the goodness of fit. Unsurprisingly, the best fit is archived by the full PHM model (Figure 2, rightmost bar). Figure 2 further shows that the min-heuristic is the most important contribution, which is suggested in Chater and Oaksford (1999). Entailment has a surprisingly small influence on the goodness of fit, while the effect of attachment is in-between. All of the heuristics perform better than the baseline goodness of fit of uniform random guessing. Thus the impact of heuristics on model performance can be ranked: G1 $>$ G3 $\gg$ G2.

[2] https://www.cc.uni-freiburg.de/code/phm

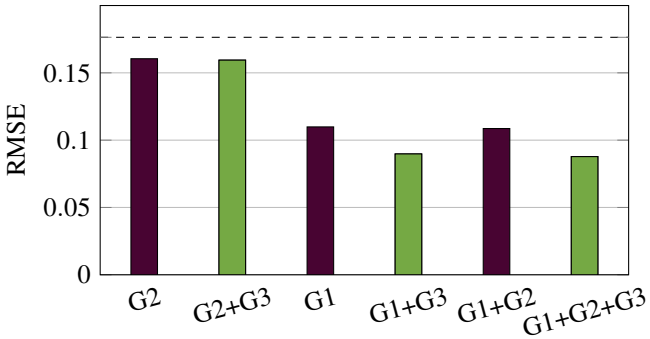

Figure 2: Impact of heuristics on the goodness of fit. The goodness of fit of the complete PHM model on the RG16 dataset is depicted by the rightmost column. When substituting the attachment heuristic G3 with randomly guessing a figure, the error increases (rightmost brown column). Using the PHM without the entailment heuristic G2 results in slightly larger errors (middle). Determining the quantifiers only via the entailment heuristic (left) results in errors close to uniform guessing (dashed line).

Falsification of the O-Heuristic. There is some incompatibility between high values for p_O (how often people accept O-conclusions for OO-syllogisms) and the O-heuristic (people avoid conclusions with an O-quantifier). All subjects using the O-heuristic implies low values for p_O. This is not the case (see Table 1), values for p_O are comparable to those of p_E and p_{ent}. Thus the O-heuristic cannot be an universally used heuristic, or a good description of aggregate data. Also, there is no evidence of an implementation of the O-heuristic by Hattori (2016) or in the predictions of the original model.

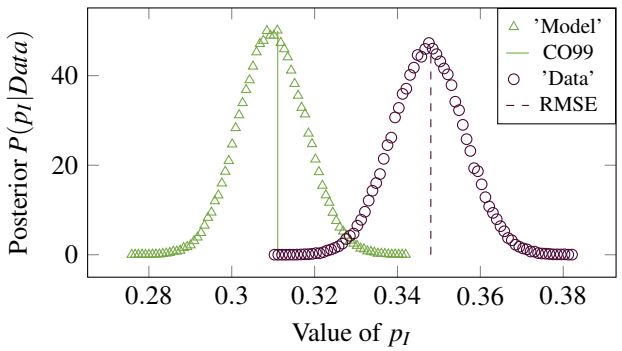

Figure 3: The p_I parameter, estimated with Bayesian methods (bell-shaped curves) or point estimates (straight lines). Neglecting the NVC answers leads to the estimates on the left hand side (Bayesian estimation from 'Model' column by Chater and Oaksford (1999) and their estimate 'CO99'). Using the full data results in the slightly different estimates on the RHS (Estimation from the 'Data' column, and a root mean squared error 'RMSE').

Bayesian modeling

To quantify the uncertainty in the parameters, we use the PyMC3-package (Salvatier, Wiecki, & Fonnesbeck, 2016) to implement a model derived from equation 14 (python code is provided[2]). The core assumption is that the categorical data of participant answers comes from a single multinomial distribution. This defines the likelihood $P(Data|\Theta)$ of every parameter Θ. We can infer posterior distributions $P(\Theta|Data)$ of the parameters from their prior distributions $P(\Theta)$ and the data aggregated over all N participants $\sum_{i=1}^{N} Data_i$:

$$P(\Theta|Data) \propto P(\sum_{i=1}^{N} Data_i|\Theta) \cdot P(\Theta) \tag{17}$$

$$P_{\Sigma}(\Theta|Data) \propto \sum_{i=1}^{N} P(Data_i|\Theta) \cdot P(\Theta) \tag{18}$$

Here, posteriors are determined up to a multiplicative factor, symbolized via the proportionality sign $\propto$. For comparison, we also infer individual posteriors from individual data (eq. 18). We chose minimally informative uniform prior distributions on the 6 parameters. The posterior distributions are estimated via Markov Chain Monte Carlo (NUTS) sampling $5 \cdot 10^5$ times with burn-in and thinning. To validate our approach, we estimate parameter posteriors (Fig. 3, left curve) using the output of the original model by Chater and Oaksford, the 'Model' data (Chater & Oaksford, 1999) and retrieve original parameter values (Fig. 3, lhs vertical line). Maxima of posteriors estimated from the 'Data' column (Chater & Oaksford, 1999) agree with parameters optimized via RMSE minimization, further validating the implementation.

Parameter Instability

Small variations in experimental conditions should not have a large effect on the cognitive processes which are described by parameters in the PHM framework. The uncertainty in parameter values due to small experiment size (here, $N = 20$) is reflected in broader posteriors, if the model assumptions are correct. One would now expect the parameters to be equal across experiments. Otherwise, the uncertainty in these parameters ought to explain the difference in the values. This is the case for some parameters in the PHM, for example the parameter p_{ent} (see Fig. 4, left hand side cluster of curves).

Table 2: Variation of best-fit values for p_I occurs under 3 different methods: RSME (eq. 15), Hattori's version of an RMSE (eq. 16) or taking the Posterior mean (eq. 17). Column names refer to experiments following Hattori (2016).

	BBJ95	JB84-3	JS78-2a	JS78-2b	RG16
RMSE	.4428	.4234	.3016	.2101	.3900
Hattori	.4288	.3817	.3144	.1690	.3993
Posterior	.4423	.4229	.2847	.1737	.3911

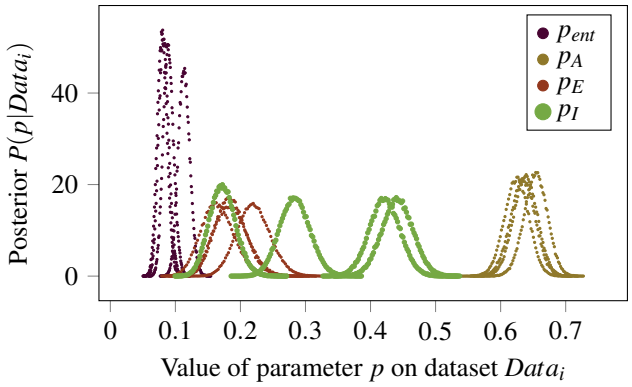

Figure 4: Variation of the PHM parameters: The posterior distributions of most parameters show appreciable overlap when calculated on the different experiments $Data_i$ collected by Hattori (2016). The posteriors of the parameter for the I quantifier (p_I, red) vary substantially. This is not plausibly explained by chance alone.

Other parameters like p_I vary more than is explained by the broadness of their posteriors. This variation is not an artifact of the bayesian parameter estimates, it occurs in comparable magnitude under three different metrics (see Table 2). It is also reported by Hattori (2016) but not explained in a satisfactory way.

Inadequacy of the aggregated-data approach. It is assumed that participant answers can be aggregated, which reduces the information content of the data, but makes modeling and data evaluation simpler. This aggregation is supposed to reduce mainly noise, not signal - which is the case if effects are distributed unimodally, but fails for multimodal distributions. Aggregating is similar to assuming every subject's mental processes come from the same distribution. To this implicit aggregation assumption, we calculate a posterior distribution for every individual. To better visualize the total distribution of the population, we then sum these posteriors (eq. 18).

Some parameters like p_{ent} seem to be unimodally distributed in the population (see Fig. 5, upper plot), although there is considerable asymmetry. These parameters also show less variation between estimates based on different experiments. For other parameters like p_I, the sum of posteriors is broad and shows multiple modes. Also, the parameter variability is more severe for these parameters. Thus, the single-distribution assumption underlying the aggregation procedure is clearly violated. This provides a qualitative explanation for the parameter instability reported by Hattori (2016).

General Discussion

Inconsistencies. Chater and Oaksford described the original implementation of the PHM based on 5 heuristics, all of them verbally specified. First, we formalize and thus fully specify this description (eq. 2 - eq. 14). We show, that this

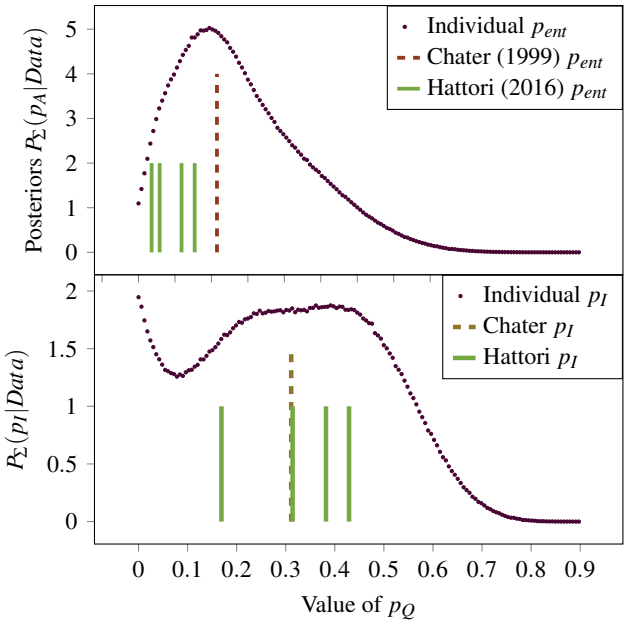

Figure 5: Top: the sum of individual posterior distributions (dark curve, compare eq. 18) shows one mode for the entailment probability p_{ent}. Bottom: for the parameter p_I, the analogous individualistic distribution (dark curve) is flatter and shows at least two modes: the peak at $p_I = 0$ and the plateau around $p_I = 0.3$. Parameter point estimates show substantially more variation in the lower plot..

formalization is functionally equivalent to the original model. We implement a bayesian model according to these formal specifications and show that it retrieves original parameter values (Fig. 3). We point out three inconsistencies within the explicit model assumptions:

1. The O-heuristic (i.e. humans tend to not accept O-type conclusions) is inconsistent with the max-heuristic (i.e. humans accept conclusions with a probability connected to the premise quantifier of maximum informativeness). Also, it is not supported by the model predictions.

2. The max-heuristic is not applied to the entailment and attachment heuristics as stated in the description. Instead, it only affects the min-heuristic.

3. The attachment heuristic is underspecified, only the cases of figure 1 and 2 are covered. An extension to figure 3 and 4 is proposed (eq. 11) and tested (Fig. 2)

Limitations. The parameters of the PHM vary across experiments, which leads to paradoxical interpretations. For example, a high p_I value can be interpreted as "in most cases, the conclusion drawn by humans will contain the min-quantifier". A low value means that this will almost never be the case. This variation is reported by Hattori (2016), but no satisfactory answer is given. Experimental conditions differing between the studies seems like an important factor, we propose another explanation. If the underlying mental processes captured by the PHM differed substantially between humans, parameters estimated from aggregates of small (N=20) studies would vary in the same way. We show that the distribution of PHM-parameters in the subject population is broad and asymmetric. It violates the implicit assumption of unimodality in some cases. Aggregating on populations of this kind results in mean values which do not reflect the actual properties of the individual reasoners. This problem has been discussed in the literature (Estes, 1956).

Potential. A great strength of the PHM is its deep theoretical motivation. The problems discussed here are of a technical nature, and can be addressed from within the PHM framework. Due to the modular nature of the PHM, contradictory heuristics (like the O-heuristic) can be omitted for more parsimony. Components that do not contribute substantial predictive power (like the entailment-heuristic) can be improved upon separately. Basing the PHM on individual instead of the currently used aggregate data might address the problem of excessive parameter variation between experiments. This approach also enables a search for groups of subjects who use only some heuristics. Thus it provides a natural way to model heuristics which appear to be contradictory on aggregate data, like e.g. the O- and max-heuristics.

Acknowledgments

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

References

Bara, B., Bucciarelli, M., & Johnson-Laird, P. (1995). Development of syllogistic reasoning. *The American journal of psychology*, 157–193.

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

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

Estes, W. (1956). The problem of inference from curves based on group data. *Psychological bulletin*, *53*(2), 134.

Hattori, M. (2016). Probabilistic representation in syllogistic reasoning: A theory to integrate mental models and heuristics. *Cognition*, *157*, 296–320.

Johnson-Laird, P., & Bara, B. (1984). Syllogistic inference. *Cognition*, *16*(1), 1–61.

Johnson-Laird, P., & Steedman, M. (1978). The psychology of syllogisms. *Cognitive psychology*, *10*(1), 64–99.

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

Salvatier, J., Wiecki, T., & Fonnesbeck, C. (2016). Probabilistic programming in python using pymc3. *PeerJ Computer Science*, *2*, e55.