Inhomogeneity in Reasoning: A Challenge for Cognitive Modeling

Lukas Elflein, Daniel Brand, Nicolas Riesterer, Marco Ragni

Human Reasoning Workshop Dresden

April 10, 2018

Introduction

- Experiments are done with individual reasoners
- Cognitive models use aggregated data
- Aggregation relies on items having similar properties
- Is human reasoning homogeneous or diverse?

Conditional Modi

Acceptance rate of Conditionals

MP: .56 - .96 $\dfrac{p \rightarrow q,\ p}{\therefore\ q}$

AC: .32 - .60 $\dfrac{p \rightarrow q,\ q}{\therefore\ p}$

MT: .35 - .65 $\dfrac{p \rightarrow q, \neg q}{\therefore\ \neg p}$

DA: .25 - .60 $\dfrac{p \rightarrow q, \neg p}{\therefore\ \neg q}$

Models for Conditionals

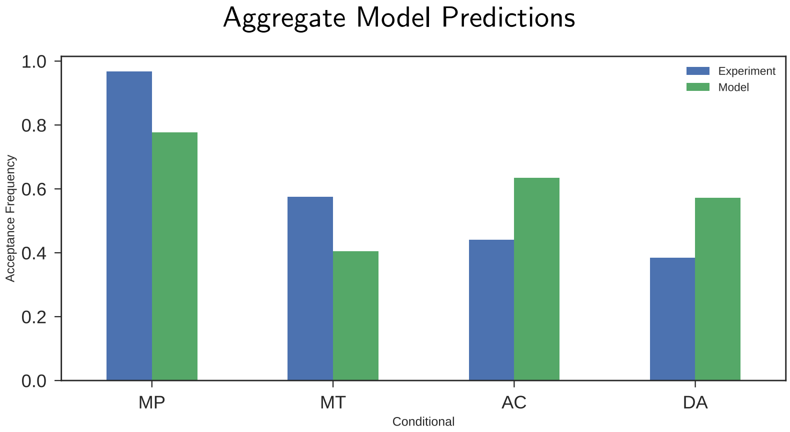
	Oaksford 2000	Dependence Model	Independence Model
MP	$1-e$	1	b
MT	$\frac{1-b-a\ e}{1-b}$	$1-a$	$1-a$
AC	$\frac{a\ (1-e)}{b}$	$\frac{a}{b}$	a
DA	$\frac{1-b-a\cdot e}{1-a}$	$1-b$	$1-b$

Bayesian Rationality models for acceptance of conditionals with
$a = P(p)$, $b = P(q)$, $e = P(\neg q|p)$

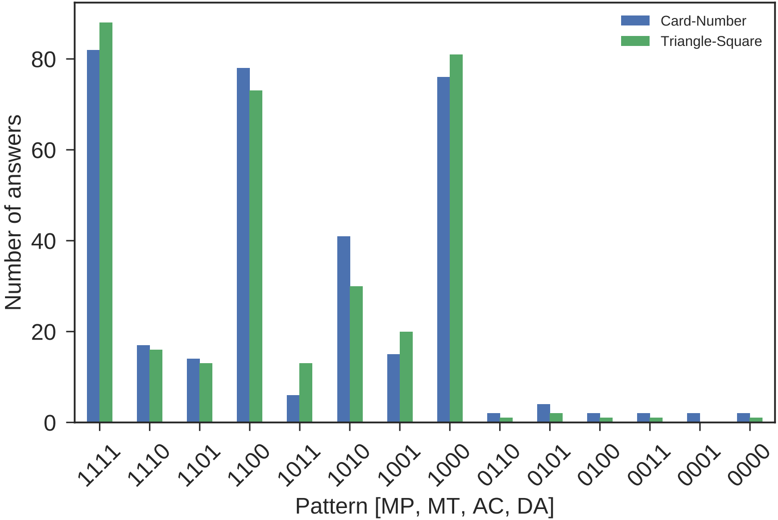
Oaksford, Chater, Larkin (2000). Probabilities and polarity biases in conditional inference.

Oaksford, Chater (1994). A rational analysis of the selection task as optimal data selection.

Aggregate Model Predictions

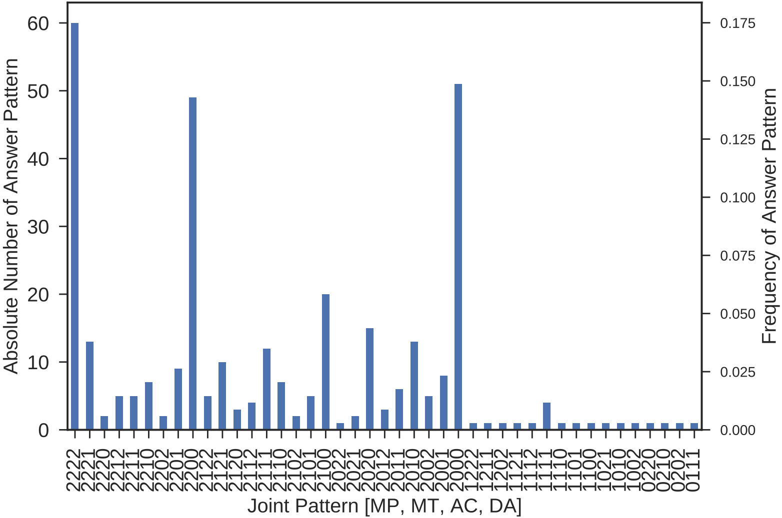


Individual Patterns



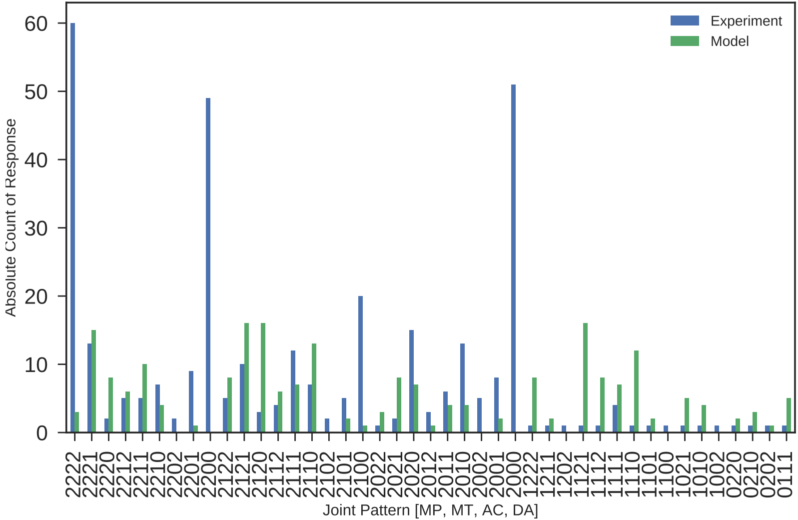
Oberauer (2006). Reasoning with conditionals: A test of formal models of four theories.

Joint Individual Patterns

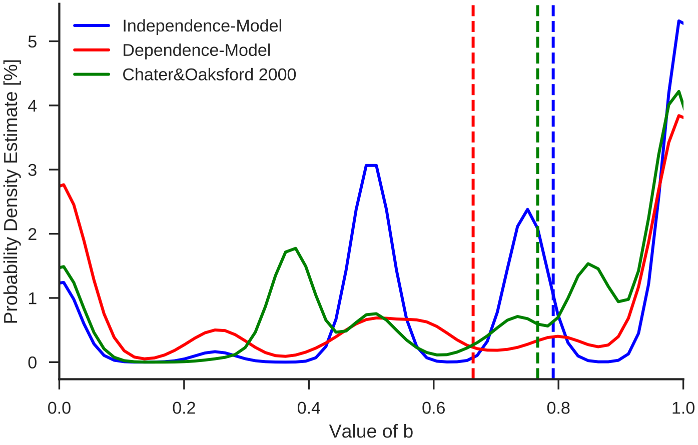


Oberauer (2006). Reasoning with conditionals: A test of formal models of four theories.

Joint Pattern Prediction with Aggregate Model



Parameter Distribution on Individual Data



Summary

- Is human reasoning homogeneous or diverse?
  Answer: There is substantial diversity in conditional reasoning.
- Aggregation masks this diversity
- What about other reasoning domains?

What are Syllogisms?

Premise 1:	All a are b
Premise 2:	Some b are c

What, if anything, follows?

Humans conclude: Some a are c.

What are Syllogisms?

Premise 1:	All a are b
Premise 2:	Some b are c

What, if anything, follows?

Humans conclude: Some a are c.
First-order logic: No Valid Conclusion

The Probability Heuristics Model

The Probability Heuristics Model (PHM)

- is a prominent probabilistic model
- models human syllogistic reasoning
- is based on 5 heuristics

Chater & Oaksford (2007). Bayesian rationality: The probabilistic approach to human reasoning.

Min-Heuristic

Premise 1:	All a are b
Premise 2:	**Some** b are c

What, if anything, follows?

The min-heuristic (G1): Choose the quantifier of the conclusion to be the same as the quantifier in the least informative premise: **Some**

$$I(\text{All}) > I(\text{Some}) > I(\text{No}) > I(\text{Some not})$$

Entailment-Heuristic

Premise 1:	All a are b
Premise 2:	**Some** b are c

What, if anything, follows?

Probabilistic entailments (G2): The next most preferred conclusion will be the entailment of the conclusion predicted by the min-heuristic: **Some not**

$$\begin{aligned} \mathit{Ent}(\text{All}) &= \text{Some}, \quad & \mathit{Ent}(\text{Some}) &= \text{Some not} \\ \mathit{Ent}(\text{Some not}) &= \text{Some}, \quad & \mathit{Ent}(\text{No}) &= \text{Some not} \end{aligned}$$

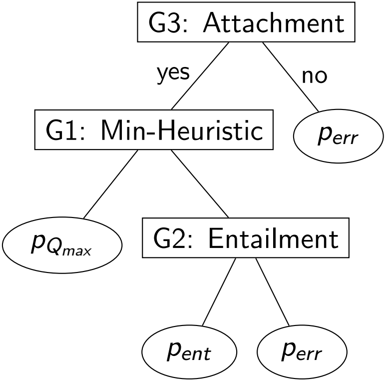
Attachment-Heuristic

Premise 1:	All **a** are b
Premise 2:	Some b are c

What, if anything, follows?

Attachment-heuristic (G3): 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: **a-c**

Example Syllogism



Premise 1:	All a are b
Premise 2:	Some b are c

What, if anything, follows?

G3: Accept a-c

G1: Accept 'Some a are c' with probability p_A

G2: Accept 'Some a are not c' with probability p_{ent}

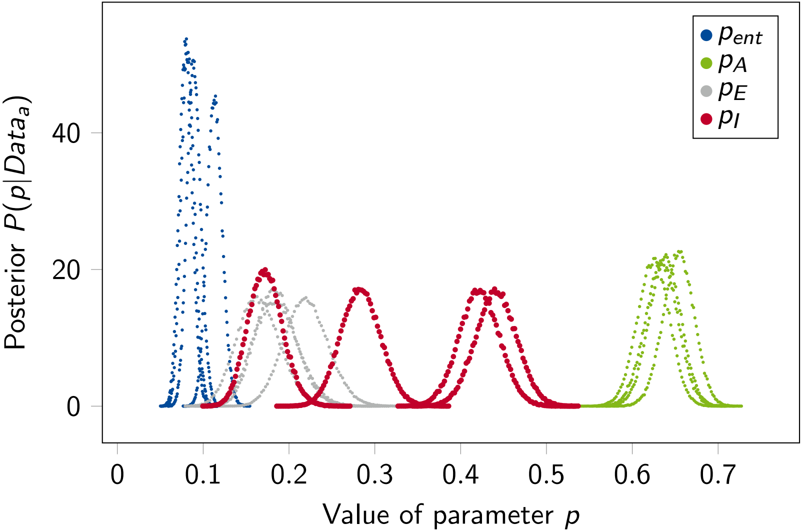
- p_A, p_I, p_E, p_O are fitted according to test heuristics T1 and T2.

Inference Methods

Approaches for aggregated data $Data_a$:

- Frequentist fits: minimize error between model estimates y_j^{mod} and experimental data $Data_j^{exp,a}$:
  $RMSE = \sqrt{\frac{1}{576}\sum_{j=1}^{576}(y_j^{mod} - Data_j^{exp,a})^2}$
- Bayesian Parameter estimate:
  $P(\Theta|Data_a) \propto P(Data_a|\Theta) \cdot P(\Theta)$

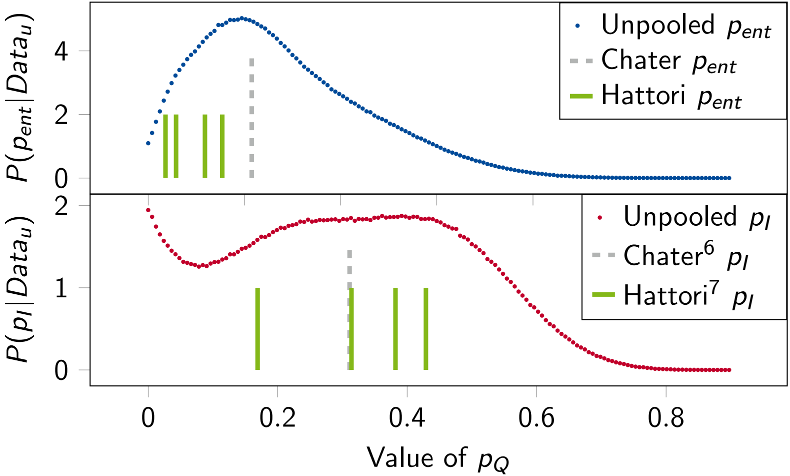
Aggregated Parameter Instability



Aggregate vs no Pooling

- Data aggregation, complete pooling:
  $P(\Theta|Data_a) \propto P(\sum_{i=1}^{N} Data_i|\Theta) \cdot P(\Theta)$
- No pooling:
  $P(\Theta|Data_u) \propto \sum_{i=1}^{N} P(Data_i|\Theta) \cdot P(\Theta)$

No Pooling vs Point Estimates



[6] Chater & Oaksford (1999). The probability heuristics model of syllogistic reasoning

[7] Hattori (2016). Probabilistic representation in syllogistic reasoning

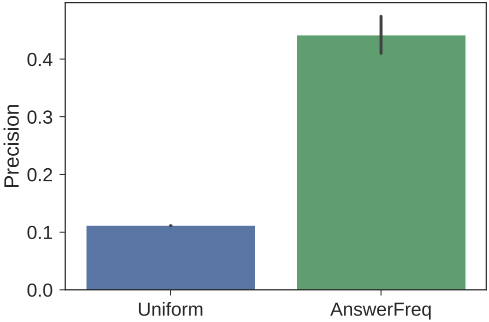
Summary

- Conditional and syllogistic reasoning is diverse in humans
- Aggregated modeling has limitations
  - Estimated parameters may vary across experiments
  - Parameters vary across participants
- Individual response pattern prediction is somewhat inaccurate

Summary

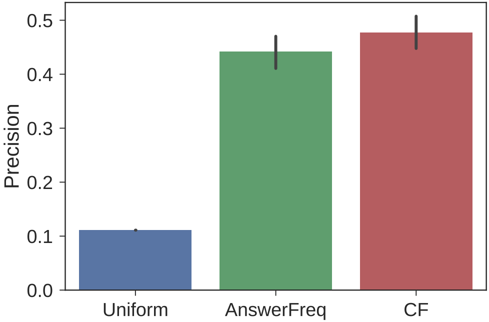
- Conditional and syllogistic reasoning is diverse in humans
- Aggregated modeling has limitations
  - Estimated parameters may vary across experiments
  - Parameters vary across participants
- Individual response pattern prediction is somewhat inaccurate
- How well do aggregate models predict individual reasoners?
- How precisely can we predict behavior in principle?
  - Is diversity in reasoning information, or is it just noise?
  - What are good predictive baselines?

Predictive Information Content



- The majority syllogistic answer (green) predicts individual answers
- Uniform guessing is substantially worse
- Can we do better than that?

Predictive Information Content

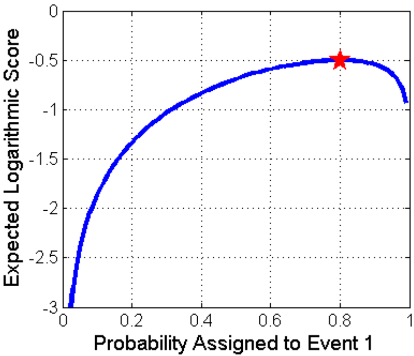


- Collaborative Filtering (red) is more accurate than the majority
- CF uses **individual information** and is **domain-agnostic**
- Useful information is lost when aggregating data

Individual Modelling Evaluation

- It is **hard to compare** different models
  - Single answer vs. multiple answers
  - Probabilities vs Ranked answers vs. unranked answers
  - 'Interpretable' models vs. 'Blackboxes'
- But we can rate their performance in a prediction task
  - Fit the model on known **training data**
  - Let it predict **test data** is has not seen
  - Rate the model performance on the test data
- Metrics for aggregated data (RMSE) mask diversity
- Which **metrics** can we use?

Proper scoring rule



- A scoring rule assigns a number to a prediction
- A proper scoring rule gives maximum score to true probability
- Properness incentivizes 'honesty'
- Improperness is unsafe for optimization ('gaming the metric')

en.wikipedia.org/wiki/Scoring_rule

Logarithmic scoring rule

$$Q = \frac{1}{64} \sum_{i=1}^{64} ln(p_i)$$

The logarithmic score

- is a proper scoring rule
- takes model probabilities p_i of the participant answer

Some models do not output probabilities!

Precision@1

$$PRC = \frac{tp}{tp + fp}$$

with:

- true positive model predictions tp
- false positives model predictions fp

For example:

- The actual participant answer is **Aac**
- The model prediction is
  - **Aac**: $PRC = 1$
  - {**Aac**, Ica}: $PRC = \frac{1}{2}$
  - {Oca, Eca}: $PRC = 0$

Precision@1 Properties

$$PRC = \frac{tp}{tp + fp}$$

Precision@1

+ is very simple
+ naturally handles incomplete models
+ does not require probabilistic predictions
- destroys the information in ranking or probabilities
- is not a proper scoring rule

Mean Reciprocal Rank (MMR)

$$MRR_p = \frac{1}{64} \sum_{a \in ans(p)} \frac{1}{rank_a}$$

with:

- the answer $ans(p) \in \{Aac, Aca, \dots\}$ given by participant p
- the rank of the model response $rank_a \in [1, 9]$

For example:

- Model prediction: **Aac** $>$ **Ica** $>$ Other
- Actual participant answer:
  - **Aac**: $MRR = 1$
  - **Iac**: $MRR = \frac{1}{2}$
  - Oca: $MRR = \frac{1}{(3+4+5+6+7+8+9)/7} = 0.17$

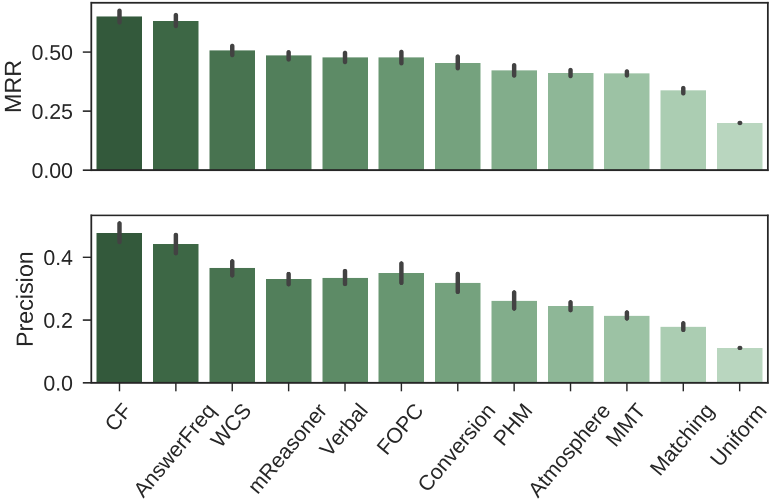
Mean Reciprocal Rank (MMR) Properties

$$MRR_p = \frac{1}{64} \sum_{a \in ans(p)} \frac{1}{rank_a}$$

The Mean Reciprocal Rank (MMR)

+ is sensitive to ranking while not requiring probabilities
+ can handle incomplete model output or missing ranks
- needs a tie-handling rule, e.g. ‘use the average rank’
- is not a proper scoring rule for simple tie-handling

Predictive Performance of Models



Summary

Results:

- Human reasoning is **diverse**
- Current models fail to capture this diversity
- Individual reasoning can be **predicted** more precisely

Open problems:

- What is a good metric for individual prediction?
- How can **cognitive theories** account for individual reasoning?
  - We need models **adapted to individual data**.
  - Can we beat Collaborative Filtering (CF)?