SEMINAR
Bayesian Theory of Mind

Lecturer: Tan Zhi Xuan
Student: Martin Nguyen (Duc Q. Nguyen)

Department of Computer Science
National University of Singapore

August 20^{th} 2025

"There is only one thing certain and that is that nothing is certain."

Gilbert K. Chesterton

Table of Contents

Table of Contents

Motivation

- People observe others' behaviours as intentional actions.

Motivation

- People observe others' behaviours as intentional actions.
- A human behaviour can be the consequence of:

Motivation

- People observe others' behaviours as intentional actions.
- A human behaviour can be the consequence of:
  - What they see/hear/touch/...

Motivation

- People observe others' behaviours as intentional actions.
- A human behaviour can be the consequence of:
  - What they see/hear/touch/...
  - What they imagine about the unobserved world

Motivation

- People observe others' behaviours as intentional actions.
- A human behaviour can be the consequence of:
  - What they see/hear/touch/...
  - What they imagine about the unobserved world
  - What they want to do

Motivation

- People observe others' behaviours as intentional actions.
- A human behaviour can be the consequence of:
  - What they see/hear/touch/...
  - What they imagine about the unobserved world
  - What they want to do

Question 1

Why do we need to understand others' behaviour (*i.e.*, infer what others want/imagine/observe)?

Motivation

Question 1

Why do we need to understand others' behaviour (*i.e.*, infer what others want/imagine/observe)?

Motivation

Question 1

Why do we need to understand others' behaviour (*i.e.*, infer what others want/imagine/observe)?

- If you understand what someone wants or imagines, you can better anticipate what they will do next.

Motivation

Question 1

Why do we need to understand others' behaviour (*i.e.*, infer what others want/imagine/observe)?

- If you understand what someone wants or imagines, you can better anticipate what they will do next.
- Human communication relies heavily on shared assumptions about others' thoughts.

Motivation

Question 1

Why do we need to understand others' behaviour (*i.e.*, infer what others want/imagine/observe)?

- If you understand what someone wants or imagines, you can better anticipate what they will do next.
- Human communication relies heavily on shared assumptions about others' thoughts.
- Understanding the reasons behind someone's actions makes it easier to empathize and respond constructively.

Motivation

Question 1

Why do we need to understand others' behaviour (*i.e.*, infer what others want/imagine/observe)?

- If you understand what someone wants or imagines, you can better anticipate what they will do next.
- Human communication relies heavily on shared assumptions about others' thoughts.
- Understanding the reasons behind someone's actions makes it easier to empathize and respond constructively.
- We learn from others' successes and mistakes by reconstructing their thought processes.

Motivation

Question 1

Why do we need to understand others' behaviour (*i.e.*, infer what others want/imagine/observe)?

- If you understand what someone wants or imagines, you can better anticipate what they will do next.
- Human communication relies heavily on shared assumptions about others' thoughts.
- Understanding the reasons behind someone's actions makes it easier to empathize and respond constructively.
- We learn from others' successes and mistakes by reconstructing their thought processes.
- The physical world is complex; people's behaviour is even more so.

Table of Contents

Preliminaries

What is the 'theory of mind'?

Preliminaries

What is the 'theory of mind'?

Theory of Mind

Theory of mind is the ability to ascribe mental states, such as beliefs, desires and intentions, to explain, predict, and justify behavior[a].

[a] Apperly and Butterfill, "Do humans have two systems to track beliefs and belief-like states?"

Preliminaries

What is the 'theory of mind'?

Theory of Mind

Theory of mind is the ability to ascribe mental states, such as beliefs, desires and intentions, to explain, predict, and justify behavior[a].

[a] Apperly and Butterfill, "Do humans have two systems to track beliefs and belief-like states?"

Now, we define the ***'mentalizing'*** process

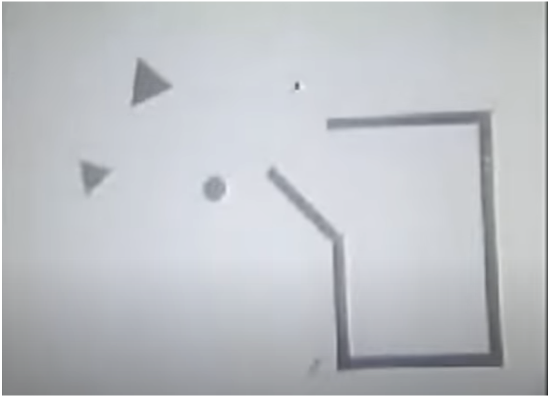
Mental State Inference

Mental state inference (or 'mentalizing') in adults is a capacity that appears in some form in infancy and persists as a richer *theory of mind* develops through the first years of life[a].

[a] Chris L. Baker et al., "Rational quantitative attribution of beliefs, desires and percepts in human mentalizing".

Examples

Heider & Simmel (1944) animation[1]



[1] Heider and Simmel, "An experimental study of apparent behavior".

Examples

Sally-Anne experiment[1]

[1] Wimmer and Perner, "Beliefs about beliefs: Representation and constraining function of wrong beliefs in young children's understanding of deception".

What mental states can we infer?

Human has many mental states:

- Percepts: What they see/hear/touch/...

What mental states can we infer?

Human has many mental states:

- Percepts: What they see/hear/touch/...
- Beliefs: What they imagine about the unobserved world

What mental states can we infer?

Human has many mental states:

- Percepts: What they see/hear/touch/...
- Beliefs: What they imagine about the unobserved world
- Desires: What they want to do

What mental states can we infer?

Human has many mental states:

- Percepts: What they see/hear/touch/...
- Beliefs: What they imagine about the unobserved world
- Desires: What they want to do
- Emotions: What they feel

What mental states can we infer?

Human has many mental states:

- Percepts: What they see/hear/touch/...
- Beliefs: What they imagine about the unobserved world
- Desires: What they want to do
- Emotions: What they feel
- . . .

What mental states can we infer?

Human has many mental states:

- Percepts: What they see/hear/touch/...
- Beliefs: What they imagine about the unobserved world
- Desires: What they want to do
- Emotions: What they feel
- ...

Core Mentalizing

- Involves observing and predicting agents' behaviors, *e.g.*, reaching for, moving toward, or manipulating objects
- Grounded in perception, action, and the physical world
- Based on line of sight and what agents can perceive
- Shaped by interactions with nearby agents who also have analogous beliefs, desires, and percepts

[a] Chris L. Baker et al., "Rational quantitative attribution of beliefs, desires and percepts in human mentalizing".

Recap

Question 2

‘’ Do you remember the types of rationality? What are the relationships with core mentalizing?

Recap

Question 2

‘’ Do you remember the types of rationality? What are the relationships with core mentalizing?

- Inferring percepts and beliefs requires understanding of epistemic rationality.
- Inferring desires requires understanding of instrumental rationality.

Recap

Question 2

‘’ Do you remember the types of rationality? What are the relationships with core mentalizing?

- Inferring percepts and beliefs requires understanding of epistemic rationality.
- Inferring desires requires understanding of instrumental rationality.

Question 3

How about the cooperative? What is the cooperation in the theory of mind?

Recap

Question 2

‘’ Do you remember the types of rationality? What are the relationships with core mentalizing?

- Inferring percepts and beliefs requires understanding of epistemic rationality.
- Inferring desires requires understanding of instrumental rationality.

Question 3

How about the cooperative? What is the cooperation in the theory of mind?

- Predicting how people will act so we can complement, not duplicate, their effort.

Recap

Question 2

" Do you remember the types of rationality? What are the relationships with core mentalizing?

- Inferring percepts and beliefs requires understanding of epistemic rationality.
- Inferring desires requires understanding of instrumental rationality.

Question 3

How about the cooperative? What is the cooperation in the theory of mind?

- Predicting how people will act so we can complement, not duplicate, their effort.
- Tailoring information to what the other person already knows or misunderstands, to maintain shared understanding.

Recap

Question 2

“ Do you remember the types of rationality? What are the relationships with core mentalizing?

- Inferring percepts and beliefs requires understanding of epistemic rationality.
- Inferring desires requires understanding of instrumental rationality.

Question 3

How about the cooperative? What is the cooperation in the theory of mind?

- Predicting how people will act so we can complement, not duplicate, their effort.
- Tailoring information to what the other person already knows or misunderstands, to maintain shared understanding.
- Changing one’s plan when detecting a mismatch between our expectations and others’ beliefs/desires.

Problem Statement

Problem

Given complete information about an agent's state and environment, and assuming an observer has access to these observations, can we develop a mathematical model for the observer's core mentalizing process (*i.e.*, their Theory of Mind)?

Table of Contents

Types of Approaches for Core Mentalizing

Based on the approach properties, they can be grouped into two types[1]:

- **Model-based[2]:** Humans have an intuitive theory of what agents think and do.
  Example: We can guess what a 6-year-old child wants, given their actions, by using some functions.

[1] Chris L. Baker et al., "Rational quantitative attribution of beliefs, desires and percepts in human mentalizing".

[2] Chris L Baker, Saxe, and Tenenbaum, "Action understanding as inverse planning"; Jern and Kemp, "A decision network account of reasoning about other people's choices"; Jara-Ettinger et al., "Children's understanding of the costs and rewards underlying rational action"; Lucas et al., "The child as econometrician: A rational model of preference understanding in children"; Oztop, Wolpert, and Kawato, "Mental state inference using visual control parameters"; Pantelis et al., "Inferring the intentional states of autonomous virtual agents".

[3] Blythe et al., *Simple heuristics that make us smart*; Zacks, "Using movement and intentions to understand simple events".

Types of Approaches for Core Mentalizing

Based on the approach properties, they can be grouped into two types[1]:

- **Model-based[2]:** Humans have an intuitive theory of what agents think and do.
  Example: We can guess what a 6-year-old child wants, given their actions, by using some functions.
- **Cue-based[3]:** Mentalizing is based on a direct mapping from low-level sensory inputs to high-level mental states via statistical associations.
  Example: You want something because you reach for it.

---

[1] Chris L. Baker et al., "Rational quantitative attribution of beliefs, desires and percepts in human mentalizing".

[2] Chris L Baker, Saxe, and Tenenbaum, "Action understanding as inverse planning"; Jern and Kemp, "A decision network account of reasoning about other people's choices"; Jara-Ettinger et al., "Children's understanding of the costs and rewards underlying rational action"; Lucas et al., "The child as econometrician: A rational model of preference understanding in children"; Oztop, Wolpert, and Kawato, "Mental state inference using visual control parameters"; Pantelis et al., "Inferring the intentional states of autonomous virtual agents".

[3] Blythe et al., *Simple heuristics that make us smart*; Zacks, "Using movement and intentions to understand simple events".

Previous Works

- Chris L Baker, Saxe, and Tenenbaum (2009) ; Jara-Ettinger *et al.* (2015) ; etc. **infer only desires** and associated notions such as goals, intentions, and preferences.
- Goodman *et al.* (2009) ; Jern and Kemp (2015) ; etc. additionally consider **inferring world states** and causal structure.
- Hawthorne-Madell and Goodman (2015) **infers beliefs** based on unobserved events.
- Butterfield *et al.* (2009) ; Shafto *et al.* (2012) jointly **infer knowledge and intentions**.

---

[1] Chris L Baker, Saxe, and Tenenbaum, "Action understanding as inverse planning".

[2] Jara-Ettinger et al., "Children's understanding of the costs and rewards underlying rational action".

[3] Goodman, Chris L Baker, and Tenenbaum, "Cause and intent: Social reasoning in causal learning".

[4] Jern and Kemp, "A decision network account of reasoning about other people's choices".

[5] Hawthorne-Madell and Goodman, "So good it has to be true: Wishful thinking in theory of mind".

[6] Butterfield et al., "Modeling aspects of theory of mind with Markov random fields".

[7] Shafto et al., "Epistemic trust: Modeling children's reasoning about others' knowledge and intent".

Do We Need to Jointly Model Core Mental States?

Do We Need to Jointly Model Core Mental States?

- Percepts (what an agent sees, hears, feels, etc.) are roots of beliefs. E.g., you see a public seminar at noon in a US institution.

Do We Need to Jointly Model Core Mental States?

- Percepts (what an agent sees, hears, feels, etc.) are roots of beliefs. E.g., you see a public seminar at noon in a US institution.
- Your beliefs and desires will lead your actions. E.g., You want a free lunch, so you go to the seminar.

Do We Need to Jointly Model Core Mental States?

- Percepts (what an agent sees, hears, feels, etc.) are roots of beliefs. E.g., you see a public seminar at noon in a US institution.
- Your beliefs and desires will lead your actions. E.g., You want a free lunch, so you go to the seminar.

Missing Element	**What Goes Wrong in Inference**
Desire	You can't tell *why* an agent chose one option over another given the same belief.
Belief	You can't explain why an agent might act "irrationally" from an outside observer's perspective (they might have false or outdated beliefs).
Percept	You can't model how beliefs arise or update in the first place, so you can't predict changes in behavior when new information appears.

How to do that?

What's wrong with prior approaches?

How to do that?

What's wrong with prior approaches?

Those approaches can not jointly rationalize percepts, beliefs, and desires as the core mentalizing requires.

How to do that?

What's wrong with prior approaches?

Those approaches can not jointly rationalize percepts, beliefs, and desires as the core mentalizing requires.

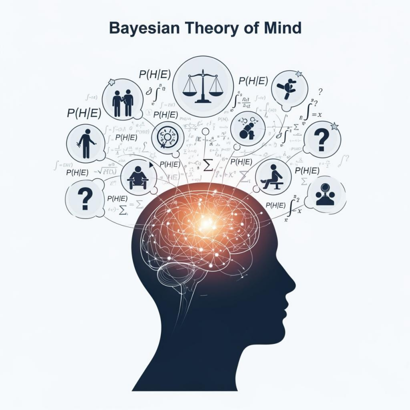


**Image generated by Imagen-4-Ultra*

Table of Contents

Overview

Bayesian Theory of Mind (BToM) contains two main components:

[1] Kaelbling, Littman, and Cassandra, "Planning and acting in partially observable stochastic domains".

Overview

Bayesian Theory of Mind (BToM) contains two main components:

- **Rational Agent Model:** A form of partially-observable Markov decision process (POMDP)[1]

[1] Kaelbling, Littman, and Cassandra, "Planning and acting in partially observable stochastic domains".

Overview

Bayesian Theory of Mind (BToM) contains two main components:

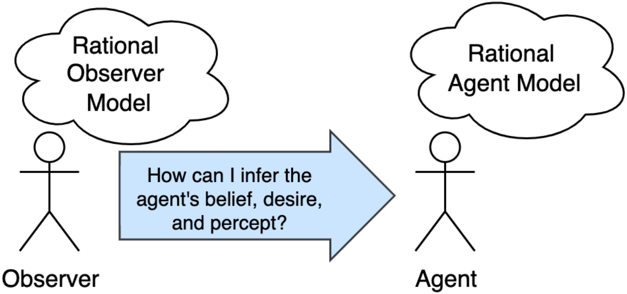
- **Rational Agent Model:** A form of partially-observable Markov decision process (POMDP)[1]
- **Rational Observer Model:** Approximate Bayesian Inference over the Rational Agent Model given necessary information

[1] Kaelbling, Littman, and Cassandra, "Planning and acting in partially observable stochastic domains".

Overview

Bayesian Theory of Mind (BToM) contains two main components:

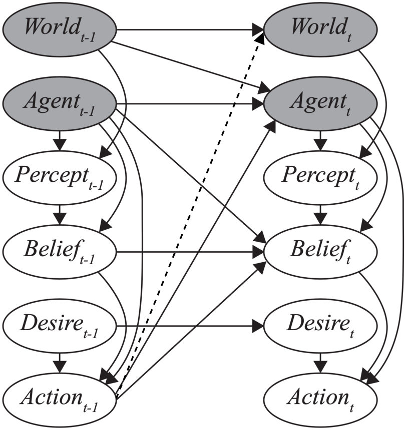
- **Rational Agent Model:** A form of partially-observable Markov decision process (POMDP)[1]
- **Rational Observer Model:** Approximate Bayesian Inference over the Rational Agent Model given necessary information

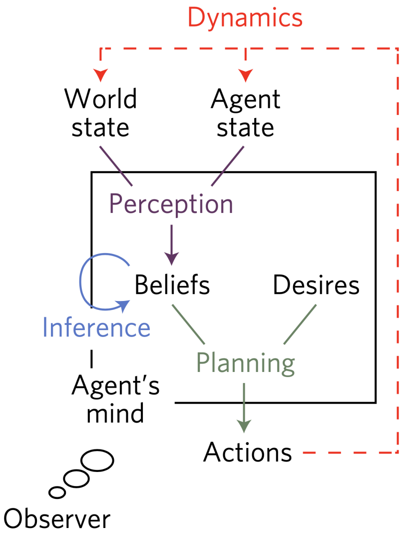


[1] Kaelbling, Littman, and Cassandra, "Planning and acting in partially observable stochastic domains".

Relations between Mental States and Environment

Theory of mind inference as a dynamic Bayes net



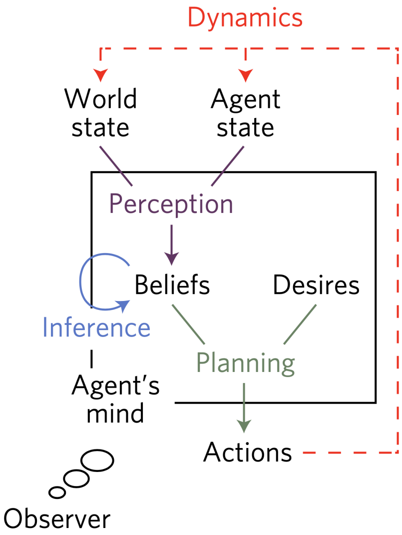


[1] Chris L. Baker et al., "Rational quantitative attribution of beliefs, desires and percepts in human mentalizing".

BToM - Rational Agent Model

We define some notations:

- $S = \langle \mathcal{X}, \mathcal{Y} \rangle$: The state space
- $x_t \in \mathcal{X}$: agent state at step t
- $y_t \in \mathcal{Y}$: world state at step t
- $o_t \in \Omega$: agent's percept at step t
- $b_t(y) = P(Y_t = y|\cdot)$: agent's belief that y is the true state at step t
- $a_t \in \mathcal{A}$: agent's action at step t
- $r(x, y, a) \in \mathcal{R}$: agent desires are represented as a reward function

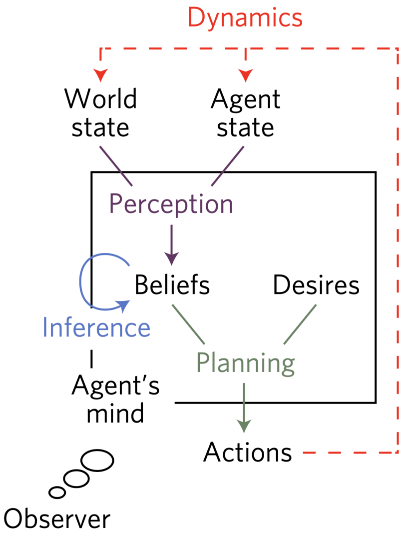


[1] Chris L. Baker et al., "Rational quantitative attribution of beliefs, desires and percepts in human mentalizing".

BToM - Rational Agent Model computation

The agent decision-making process is modeled given initial belief b_0 and desire r as follows.

$$
\begin{aligned}
x_t &\sim P(x_t | x_{t-1}, y_{t-1}, a_{t-1}) \\
y_t &\sim P(y_t | y_{t-1}, a_{t-1}) \\
o_t &\sim P(o_t | x_t, y_t) \\
b_t &\sim P(b_t | b_{t-1}, o_t) \\
a_t &\sim P(a_t | b_t, x_t, r)
\end{aligned} \tag{1}
$$



[1] Chris L. Baker et al., "Rational quantitative attribution of beliefs, desires and percepts in human mentalizing".

BToM - Rational Agent Model computation

We can abstract the computation into two steps[1]

[1] Chris L. Baker et al., "Rational quantitative attribution of beliefs, desires and percepts in human mentalizing".

[2] Lovejoy, "Computationally feasible bounds for partially observed Markov decision processes".

[3] Kurniawati, Hsu, and Lee, "Sarsop: Efficient point-based pomdp planning by approximating optimally reachable belief spaces."

BToM - Rational Agent Model computation

We can abstract the computation into two steps[1]

- **Belief Update:** $b_t = BU(o_t, x_t, x_{t-1}, y_t, y_{t-1}, a_{t-1}, b_{t-1})$, where $b_t(y) \propto P(o_t|x_t, y_t)P(x_t|x_{t-1}, y_{t-1}, a_{t-1})P(y_t|y_{t-1}, a_{t-1})b_{t-1}(y)$.

[1] Chris L. Baker et al., "Rational quantitative attribution of beliefs, desires and percepts in human mentalizing".

[2] Lovejoy, "Computationally feasible bounds for partially observed Markov decision processes".

[3] Kurniawati, Hsu, and Lee, "Sarsop: Efficient point-based pomdp planning by approximating optimally reachable belief spaces."

BToM - Rational Agent Model computation

We can abstract the computation into two steps[1]

- **Belief Update:** $b_t = BU(o_t, x_t, x_{t-1}, y_t, y_{t-1}, a_{t-1}, b_{t-1})$, where $b_t(y) \propto P(o_t|x_t, y_t)P(x_t|x_{t-1}, y_{t-1}, a_{t-1})P(y_t|y_{t-1}, a_{t-1})b_{t-1}(y)$.
- **Planning:** $a_t \sim P(a_t|b_t, x_t, r)$. There are multiple planning algorithms; Baker *et al.* used a grid-based value iteration algorithm[2] and the SARSOP algorithm[3].

[1] Chris L. Baker et al., "Rational quantitative attribution of beliefs, desires and percepts in human mentalizing".

[2] Lovejoy, "Computationally feasible bounds for partially observed Markov decision processes".

[3] Kurniawati, Hsu, and Lee, "Sarsop: Efficient point-based pomdp planning by approximating optimally reachable belief spaces."

BToM - Rational Agent Model computation

We can abstract the computation into two steps[1]

- **Belief Update:** $b_t = BU(o_t, x_t, x_{t-1}, y_t, y_{t-1}, a_{t-1}, b_{t-1})$, where $b_t(y) \propto P(o_t|x_t, y_t)P(x_t|x_{t-1}, y_{t-1}, a_{t-1})P(y_t|y_{t-1}, a_{t-1})b_{t-1}(y)$.
- **Planning:** $a_t \sim P(a_t|b_t, x_t, r)$. There are multiple planning algorithms; Baker *et al.* used a grid-based value iteration algorithm[2] and the SARSOP algorithm[3].

Question 4

The BToM model uses a POMDP solver to compute what actions a_t a rational agent should do given their current belief b_t and desire r. However, POMDP solvers typically result in plans/policies that are deterministic (*i.e.*, a single optimal action is taken at each belief state). How does the BToM model turn this into a distribution over actions instead, $P(a_t|b_t, r)$? Is this a reasonable probabilistic model of how agents select actions?

[1] Chris L. Baker et al., "Rational quantitative attribution of beliefs, desires and percepts in human mentalizing".

[2] Lovejoy, "Computationally feasible bounds for partially observed Markov decision processes".

[3] Kurniawati, Hsu, and Lee, "Sarsop: Efficient point-based pomdp planning by approximating optimally reachable belief spaces."

BToM - Rational Observer Model

The inference process of the observer can be conceptualized as follows.

BToM - Rational Observer Model

The inference process of the observer can be conceptualized as follows.

- At the beginning, observers do not know the agent's belief, desire, or percept.

BToM - Rational Observer Model

The inference process of the observer can be conceptualized as follows.

- At the beginning, observers do not know the agent's belief, desire, or percept.
- They start with initial assumptions about the agent's states (*e.g.*, assigning equal probability to all possible beliefs).

BToM - Rational Observer Model

The inference process of the observer can be conceptualized as follows.

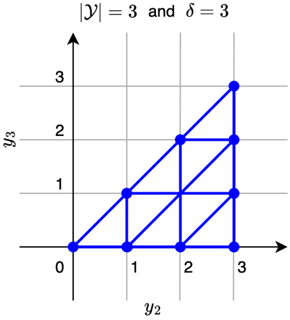
- At the beginning, observers do not know the agent's belief, desire, or percept.
- They start with initial assumptions about the agent's states (*e.g.*, assigning equal probability to all possible beliefs).
- As they observe the agent's trajectory (*i.e.*, world states and agent states), they update these assumptions so that the inferred beliefs/desires/percepts are consistent with the observed information.

Belief and Desire Priors

Belief space:

$$\Delta^{|\mathcal{Y}|-1} = \left\{ p \in \mathbb{R}^{|Y|} : p_i \geq 0, \sum_{i=1}^{|Y|} p_i = 1 \right\}$$

- Discretize using Freudenthal Triangulation with resolution δ
- Number of belief points b_0^i: $m(0) = \binom{|Y|-1+\delta}{|Y|-1}$.



[1] Chris L. Baker et al., "Rational quantitative attribution of beliefs, desires and percepts in human mentalizing".

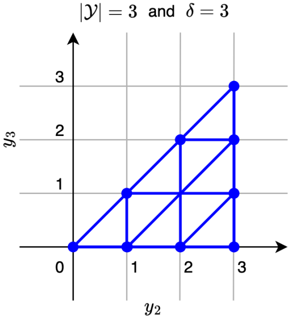
[2] Lovejoy, "Computationally feasible bounds for partially observed Markov decision processes".

Belief and Desire Priors

Belief space:

$$\Delta^{|\mathcal{Y}|-1} = \left\{ p \in \mathbb{R}^{|Y|} : p_i \geq 0, \sum_{i=1}^{|Y|} p_i = 1 \right\}$$

- Discretize using Freudenthal Triangulation with resolution δ
- Number of belief points b_0^i: $m(0) = \binom{|Y|-1+\delta}{|Y|-1}$.



Desire (reward) space:

- For each goal $g \in \mathcal{G}$, discretize reward values with resolution η
- Number of reward functions r^k: $n = \eta^{|\mathcal{G}|}$.

[1] Chris L. Baker et al., "Rational quantitative attribution of beliefs, desires and percepts in human mentalizing".
[2] Lovejoy, "Computationally feasible bounds for partially observed Markov decision processes".

BToM - Rational Observer Model

Given the agent's trajectory up to step T, we infer beliefs and desires by

$$P(b^i_{0:T}, r^k_{\mathcal{G}} | x_{1:T}, y_{1:T}) \propto P(b^i_{1:T}, r^k_{\mathcal{G}}, x_{1:T}, y_{1:T})$$

[1] Chris L. Baker et al., "Rational quantitative attribution of beliefs, desires and percepts in human mentalizing".

BToM - Rational Observer Model

Given the agent's trajectory up to step T, we infer beliefs and desires by

$$\begin{aligned} P(b^i_{0:T}, r^k_{\mathcal{G}} | x_{1:T}, y_{1:T}) &\propto P(b^i_{1:T}, r^k_{\mathcal{G}}, x_{1:T}, y_{1:T}) \\ &\propto P(b^i_0, r^k_{\mathcal{G}}) \prod_{t=1}^{T} P(b^i_t, x_t, y_t | b^i_{t-1}, r^k_{\mathcal{G}}, x_{t-1}, y_{t-1}) \end{aligned} \tag{2}$$

[1] Chris L. Baker et al., "Rational quantitative attribution of beliefs, desires and percepts in human mentalizing".

BToM - Rational Observer Model

Given the agent's trajectory up to step T, we infer beliefs and desires by

$$\begin{aligned} P(b^i_{0:T}, r^k_{\mathcal{G}} | x_{1:T}, y_{1:T}) &\propto P(b^i_{1:T}, r^k_{\mathcal{G}}, x_{1:T}, y_{1:T}) \\ &\propto P(b^i_0, r^k_{\mathcal{G}}) \prod_{t=1}^{T} P(b^i_t, x_t, y_t | b^i_{t-1}, r^k_{\mathcal{G}}, x_{t-1}, y_{t-1}) \end{aligned} \tag{2}$$

With $s_t = \langle x_t, y_t \rangle$ and $P(b_t | s_t, o_t) = P(b_t | o_t)$, we have

$$P(b^i_{0:T}, r^k_{\mathcal{G}} \mid s_{1:T}) \propto P(b^i_0, r^k_{\mathcal{G}}) \prod_{t=1}^{T} P(b^i_t, s_t \mid b^i_{t-1}, s_{t-1}, r^k_{\mathcal{G}})$$

BToM - Rational Observer Model

Given the agent's trajectory up to step T, we infer beliefs and desires by

$$\begin{aligned} P(b^i_{0:T}, r^k_{\mathcal{G}} | x_{1:T}, y_{1:T}) &\propto P(b^i_{1:T}, r^k_{\mathcal{G}}, x_{1:T}, y_{1:T}) \\ &\propto P(b^i_0, r^k_{\mathcal{G}}) \prod_{t=1}^{T} P(b^i_t, x_t, y_t | b^i_{t-1}, r^k_{\mathcal{G}}, x_{t-1}, y_{t-1}) \end{aligned} \tag{2}$$

With $s_t = \langle x_t, y_t \rangle$ and $P(b_t | s_t, o_t) = P(b_t | o_t)$, we have

$$\begin{aligned} P(b^i_{0:T}, r^k_{\mathcal{G}} \mid s_{1:T}) &\propto P(b^i_0, r^k_{\mathcal{G}}) \prod_{t=1}^{T} P(b^i_t, s_t \mid b^i_{t-1}, s_{t-1}, r^k_{\mathcal{G}}) \\ &= P(b^i_0, r^k_{\mathcal{G}}) \prod_{t=1}^{T} \sum_{o_t} P(b^i_t, s_t \mid b^i_{t-1}, s_{t-1}, r^k_{\mathcal{G}}, o_t) P(o_t | b^i_{t-1}, s_{t-1}, r^k_{\mathcal{G}}) \end{aligned}$$

BToM - Rational Observer Model

Given the agent's trajectory up to step T, we infer beliefs and desires by

$$
\begin{aligned}
P(b^i_{0:T}, r^k_{\mathcal{G}} | x_{1:T}, y_{1:T}) &\propto P(b^i_{1:T}, r^k_{\mathcal{G}}, x_{1:T}, y_{1:T}) \\
&\propto P(b^i_0, r^k_{\mathcal{G}}) \prod_{t=1}^{T} P(b^i_t, x_t, y_t | b^i_{t-1}, r^k_{\mathcal{G}}, x_{t-1}, y_{t-1})
\end{aligned}
\tag{2}
$$

With $s_t = \langle x_t, y_t \rangle$ and $P(b_t|s_t, o_t) = P(b_t|o_t)$, we have

$$
\begin{aligned}
P(b^i_{0:T}, r^k_{\mathcal{G}} \mid s_{1:T}) &\propto P(b^i_0, r^k_{\mathcal{G}}) \prod_{t=1}^{T} P(b^i_t, s_t \mid b^i_{t-1}, s_{t-1}, r^k_{\mathcal{G}}) \\
&= P(b^i_0, r^k_{\mathcal{G}}) \prod_{t=1}^{T} \sum_{o_t} P(b^i_t, s_t \mid b^i_{t-1}, s_{t-1}, r^k_{\mathcal{G}}, o_t) P(o_t | b^i_{t-1}, s_{t-1}, r^k_{\mathcal{G}}) \\
&= P(b^i_0, r^k_{\mathcal{G}}) \prod_{t=1}^{T} \sum_{o_t} P(b^i_t \mid o_t, b^i_{t-1}) P(s_t \mid b^i_{t-1}, s_{t-1}, r^k_{\mathcal{G}}, o_t) P(o_t | b^i_{t-1}, s_{t-1}, r^k_{\mathcal{G}})
\end{aligned}
$$

[1] Chris L. Baker et al., "Rational quantitative attribution of beliefs, desires and percepts in human mentalizing".

BToM - Rational Observer Model

$$P(b^i_{0:T}, r^k_{\mathcal{G}} \mid s_{1:T}) \propto P(b^i_0, r^k_{\mathcal{G}}) \prod_{t=1}^{T} \sum_{o_t} P(b^i_t \mid o_t, b^i_{t-1}) P(s_t, o_t \mid b^i_{t-1}, s_{t-1}, r^k_{\mathcal{G}})$$

BToM - Rational Observer Model

$$
\begin{aligned}
P(b^i_{0:T}, r^k_{\mathcal{G}} \mid s_{1:T}) &\propto P(b^i_0, r^k_{\mathcal{G}}) \prod_{t=1}^{T} \sum_{o_t} P(b^i_t \mid o_t, b^i_{t-1}) P(s_t, o_t \mid b^i_{t-1}, s_{t-1}, r^k_{\mathcal{G}}) \\
&= P(b^i_0, r^k_{\mathcal{G}}) \prod_{t=1}^{T} \sum_{o_t} P(b^i_t \mid o_t, b^i_{t-1}) P(o_t|s_t) P(s_t \mid b^i_{t-1}, s_{t-1}, r^k_{\mathcal{G}})
\end{aligned}
$$

BToM - Rational Observer Model

$$
\begin{aligned}
P(b^i_{0:T}, r^k_{\mathcal{G}} \mid s_{1:T}) &\propto P(b^i_0, r^k_{\mathcal{G}}) \prod_{t=1}^{T} \sum_{o_t} P(b^i_t \mid o_t, b^i_{t-1}) P(s_t, o_t \mid b^i_{t-1}, s_{t-1}, r^k_{\mathcal{G}}) \\
&= P(b^i_0, r^k_{\mathcal{G}}) \prod_{t=1}^{T} \sum_{o_t} P(b^i_t \mid o_t, b^i_{t-1}) P(o_t | s_t) P(s_t \mid b^i_{t-1}, s_{t-1}, r^k_{\mathcal{G}}) \\
&= P(b^i_0, r^k_{\mathcal{G}}) \prod_{t=1}^{T} \Bigg(\sum_{o_t} P(b^i_t \mid o_t, b^i_{t-1}) P(o_t | s_t) \cdot \\
&\qquad \sum_{a_{t-1}} P(s_t \mid b^i_{t-1}, s_{t-1}, r^k_{\mathcal{G}}, a_{t-1}) P(a_{t-1} \mid b^i_{t-1}, r^k_{\mathcal{G}}) \Bigg)
\end{aligned}
$$

BToM - Rational Observer Model

$$
\begin{aligned}
P(b^i_{0:T}, r^k_{\mathcal{G}} \mid s_{1:T}) &\propto P(b^i_0, r^k_{\mathcal{G}}) \prod_{t=1}^{T} \sum_{o_t} P(b^i_t \mid o_t, b^i_{t-1}) P(s_t, o_t \mid b^i_{t-1}, s_{t-1}, r^k_{\mathcal{G}}) \\
&= P(b^i_0, r^k_{\mathcal{G}}) \prod_{t=1}^{T} \sum_{o_t} P(b^i_t \mid o_t, b^i_{t-1}) P(o_t|s_t) P(s_t \mid b^i_{t-1}, s_{t-1}, r^k_{\mathcal{G}}) \\
&= P(b^i_0, r^k_{\mathcal{G}}) \prod_{t=1}^{T} \Bigg(\sum_{o_t} P(b^i_t \mid o_t, b^i_{t-1}) P(o_t|s_t) \cdot \\
&\qquad \sum_{a_{t-1}} P(s_t \mid b^i_{t-1}, s_{t-1}, r^k_{\mathcal{G}}, a_{t-1}) P(a_{t-1} \mid b^i_{t-1}, r^k_{\mathcal{G}}) \Bigg) \\
&= P(b^i_0, r^k_{\mathcal{G}}) \prod_{t=1}^{T} \Bigg(\sum_{o_t} P(b^i_t \mid o_t, b^i_{t-1}) P(o_t|s_t) \cdot \\
&\qquad \sum_{a_{t-1}} P(s_t \mid s_{t-1}, a_{t-1}) P(a_{t-1} \mid b^i_{t-1}, r^k_{\mathcal{G}}) \Bigg)
\end{aligned}
$$

[1] Chris L. Baker et al., "Rational quantitative attribution of beliefs, desires and percepts in human mentalizing".

BToM - Rational Observer Model

Analogously, the observer can perform percept inference given only the agent's trajectory by

$$
\begin{aligned}
P(y|x_{1:T}) &= \sum_{b_t^i, r_{\mathcal{G}}^k} P(b_t^i, r_{\mathcal{G}}^k, y|x_{1:T}) \\
&= \sum_{b_t^i, r_{\mathcal{G}}^k} P(b_t^i, r_{\mathcal{G}}^k|x_{1:T}, y)P(y).
\end{aligned}
\tag{3}
$$

As $P(b_t^i, r_{\mathcal{G}}^k|x_{1:T}, y)$ has been computed before and $P(y)$ can be computed from environment, the $P(y|x_{1:T})$ is computable.
**In this study, we consider the case where the world states y_t remain fixed. Thus, $y = y_1 = \cdots = y_T$.*

[1] Chris L. Baker et al., "Rational quantitative attribution of beliefs, desires and percepts in human mentalizing".

Table of Contents

Experiment Setup

To validate the proposed model, the authors perform two experiments:

- **Experiment 1:** Participants saw a large number of dynamic scenarios and made quantitative inferences about agents' *beliefs* and *desires* given their observable actions.

[1] Chris L. Baker et al., "Rational quantitative attribution of beliefs, desires and percepts in human mentalizing".

Experiment Setup

To validate the proposed model, the authors perform two experiments:

- **Experiment 1:** Participants saw a large number of dynamic scenarios and made quantitative inferences about agents' *beliefs* and *desires* given their observable actions.
- **Experiment 2:** Participants made inferences about agents' *percepts* and aspects of the world that only the agent could perceive.

[1] Chris L. Baker et al., "Rational quantitative attribution of beliefs, desires and percepts in human mentalizing".

Experiment Setup

To validate the proposed model, the authors perform two experiments:

- **Experiment 1:** Participants saw a large number of dynamic scenarios and made quantitative inferences about agents' *beliefs* and *desires* given their observable actions.
- **Experiment 2:** Participants made inferences about agents' *percepts* and aspects of the world that only the agent could perceive.

Both experiments were tested using bootstrap cross-validated (BSCV) correlations with disjoint training and testing sets.

[1] Chris L. Baker et al., "Rational quantitative attribution of beliefs, desires and percepts in human mentalizing".

Baselines

The authors compare the proposed model with three others:

- **TrueBelief:** (*model-based*) Similar to BToM, but the agent knows the true world state (*i.e.* its belief matches the real world states).

[1] Chris L. Baker et al., "Rational quantitative attribution of beliefs, desires and percepts in human mentalizing".

Baselines

The authors compare the proposed model with three others:

- **TrueBelief:** (*model-based*) Similar to BToM, but the agent knows the true world state (*i.e.* its belief matches the real world states).
- **NoCost:** (*model-based*) Similar to BToM, the agents plan their actions without optimizing cost.

[1] Chris L. Baker et al., "Rational quantitative attribution of beliefs, desires and percepts in human mentalizing".

Baselines

The authors compare the proposed model with three others:

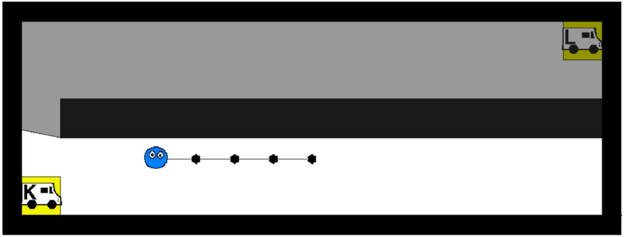
- **TrueBelief:** (*model-based*) Similar to BToM, but the agent knows the true world state (*i.e.* its belief matches the real world states).
- **NoCost:** (*model-based*) Similar to BToM, the agents plan their actions without optimizing cost.
- **MotionHeuristic:** (*cue-based*) This model maps cues extracted from the agent's motion and environment directly onto the observer's judgements of agents' beliefs, desires, and percepts of the world.

[1] Chris L. Baker et al., "Rational quantitative attribution of beliefs, desires and percepts in human mentalizing".

Experiment 1: Food Trucks

- Food trucks: Korean (K), Lebanese (L), and Mexican (M).
- There are only two parking slots for the trucks on campus.
- At least one truck parks in the lower-left every day.
- The agent is a student going to lunch with unknown truck preference.

The observer will be given the agent's path. The observer is required to rate the agent's *truck preference* and the agent's initial belief about the *possible occupant* of the far parking spot.



[1] Chris L. Baker et al., "Rational quantitative attribution of beliefs, desires and percepts in human mentalizing".

Experiment 1: Food Trucks

- Food trucks: Korean (K), Lebanese (L), and Mexican (M).
- There are only two parking slots for the trucks on campus.
- At least one truck parks in the lower-left every day.
- The agent is a student going to lunch with unknown truck preference.

The observer will be given the agent's path. The observer is required to rate the agent's *truck preference* and the agent's initial belief about the *possible occupant* of the far parking spot.

[1] Chris L. Baker et al., "Rational quantitative attribution of beliefs, desires and percepts in human mentalizing".

Experiment 1: Food Trucks

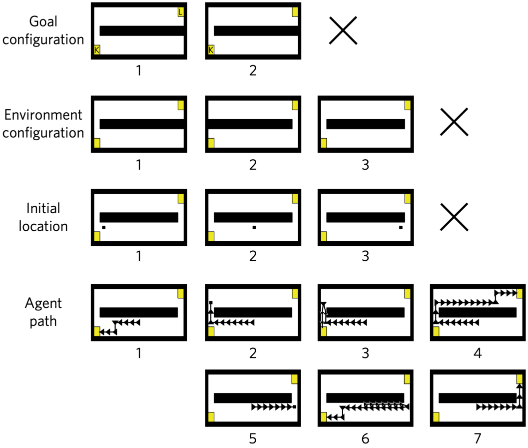
- Food trucks: Korean (K), Lebanese (L), and Mexican (M).
- There are only two parking slots for the trucks on campus.
- At least one truck parks in the lower-left every day.
- The agent is a student going to lunch with unknown truck preference.

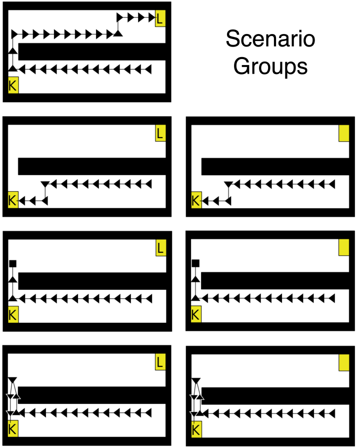
The observer will be given the agent's path. The observer is required to rate the agent's *truck preference* and the agent's initial belief about the *possible occupant* of the far parking spot.

[1] Chris L. Baker et al., "Rational quantitative attribution of beliefs, desires and percepts in human mentalizing".

Experiment 1: Food Trucks

The experiment varies 4 factors and groups scenarios into 7 sets.





[1] Chris L. Baker et al., "Rational quantitative attribution of beliefs, desires and percepts in human mentalizing".

Experiment 1 Results

BSCV used 100,000 iterations, with training folds containing 4/7 of scenario types, and testing folds containing 3/7 of scenario types. Values reported are median BSCV correlations with 95% confidence intervals. (*) indicates r-values which are significantly less than those of BToM ($p < 0.00001$).

Table 1: BSCV analysis of model predictions for individual scenarios

r (BSCV)	**BToM**	**TimeBelief**	**NoCost**	**MotionHeuristic**
Desire (individual)	**0.91 (0.89, 0.92)**	0.72 (0.68, 0.77)*	0.75 (0.69, 0.81)*	0.62 (0.51, 0.70)*
Belief (individual)	**0.78 (0.72, 0.85)**	-0.02 (-0.16, 0.11)*	0.10 (0.05, 0.15)*	0.79 (0.71, 0.84)

Table 2: BSCV analysis of model predictions for grouped scenarios

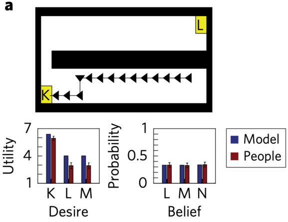
r (BSCV)	**BToM**	**TimeBelief**	**NoCost**	**MotionHeuristic**
Desire (grouped)	**0.97 (0.95, 0.98)**	0.78 (0.70, 0.86)*	0.80 (0.39, 0.96)*	0.65 (-0.09, 0.87)*
Belief (grouped)	**0.91 (0.87, 0.98)**	-0.04 (-0.52, 0.49)*	0.19 (0.02, 0.93)	0.77 (0.31, 0.93)

[1] Chris L. Baker et al., "Rational quantitative attribution of beliefs, desires and percepts in human mentalizing".

Experiment 1 Results (cont.)

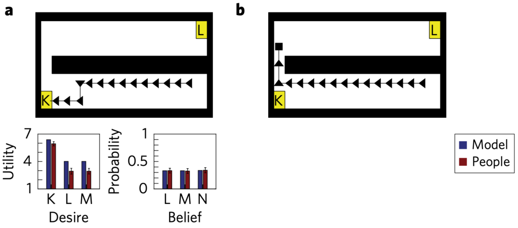
Comparing BToM and mean human ($n = 16$) desire and belief inferences from seven key scenario types.

a



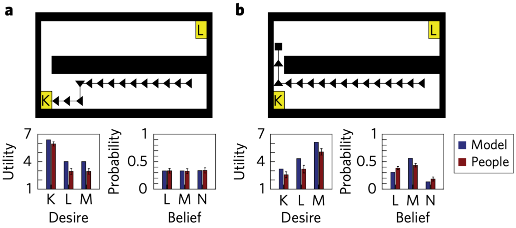
Experiment 1 Results (cont.)

Comparing BToM and mean human ($n = 16$) desire and belief inferences from seven key scenario types.



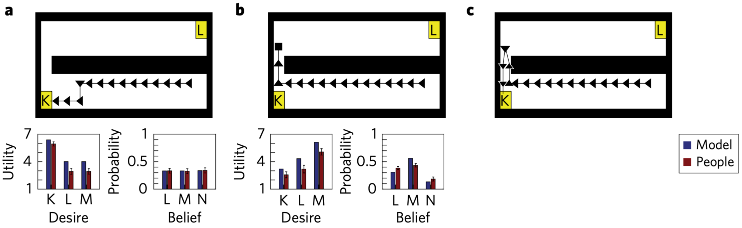
Experiment 1 Results (cont.)

Comparing BToM and mean human ($n = 16$) desire and belief inferences from seven key scenario types.



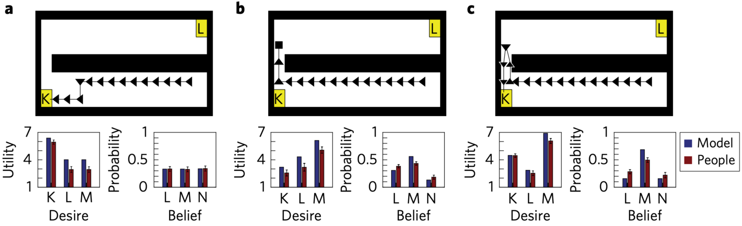
Experiment 1 Results (cont.)

Comparing BToM and mean human ($n = 16$) desire and belief inferences from seven key scenario types.



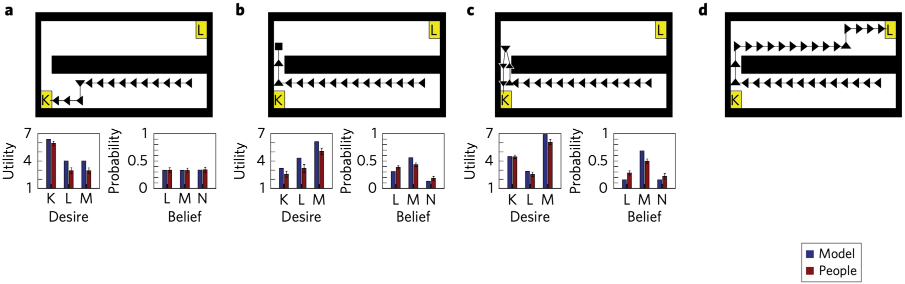
Experiment 1 Results (cont.)

Comparing BToM and mean human ($n = 16$) desire and belief inferences from seven key scenario types.



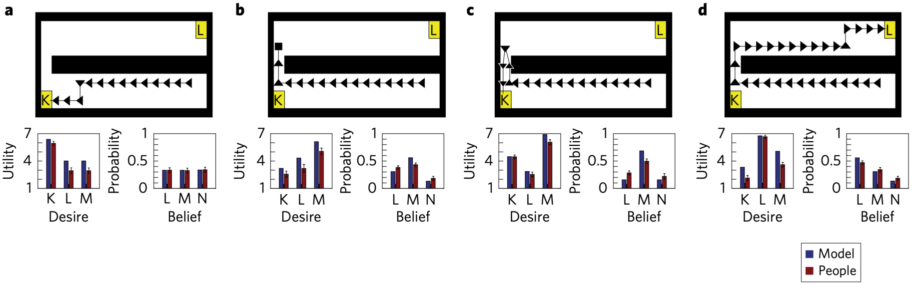
Experiment 1 Results (cont.)

Comparing BToM and mean human ($n = 16$) desire and belief inferences from seven key scenario types.



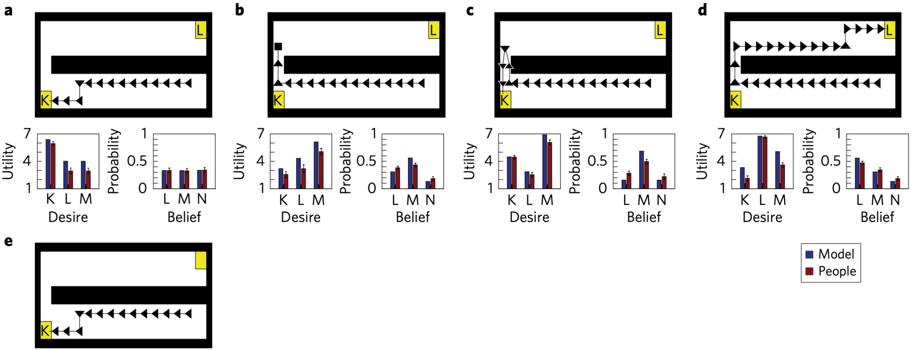
Experiment 1 Results (cont.)

Comparing BToM and mean human ($n = 16$) desire and belief inferences from seven key scenario types.



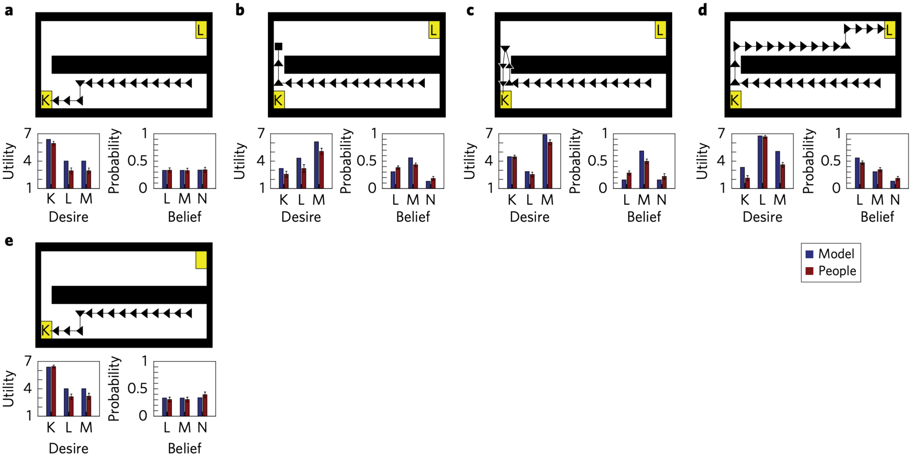
Experiment 1 Results (cont.)

Comparing BToM and mean human ($n = 16$) desire and belief inferences from seven key scenario types.



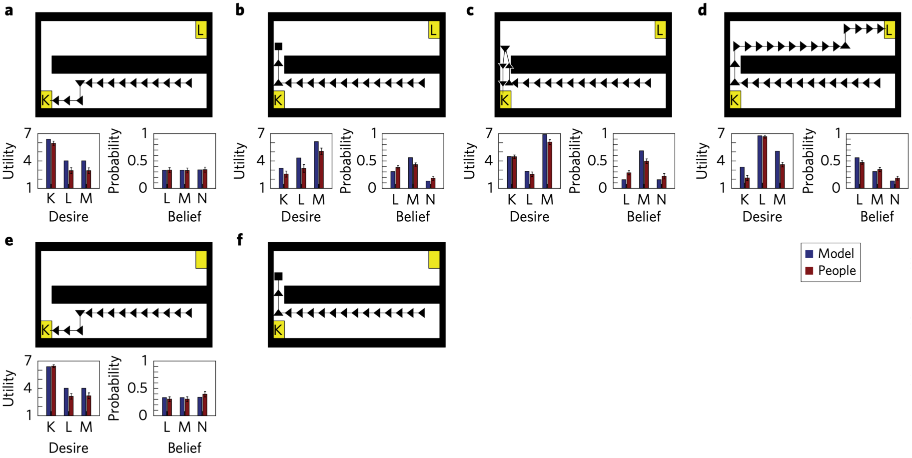
Experiment 1 Results (cont.)

Comparing BToM and mean human ($n = 16$) desire and belief inferences from seven key scenario types.



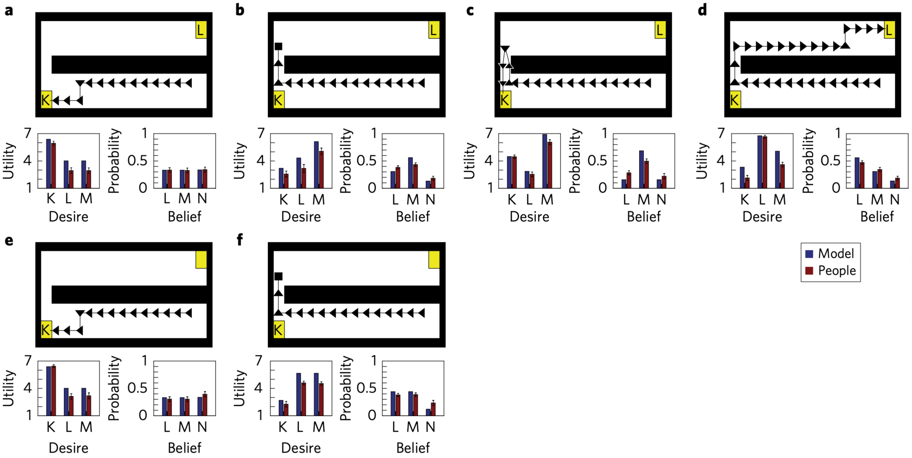
Experiment 1 Results (cont.)

Comparing BToM and mean human ($n = 16$) desire and belief inferences from seven key scenario types.



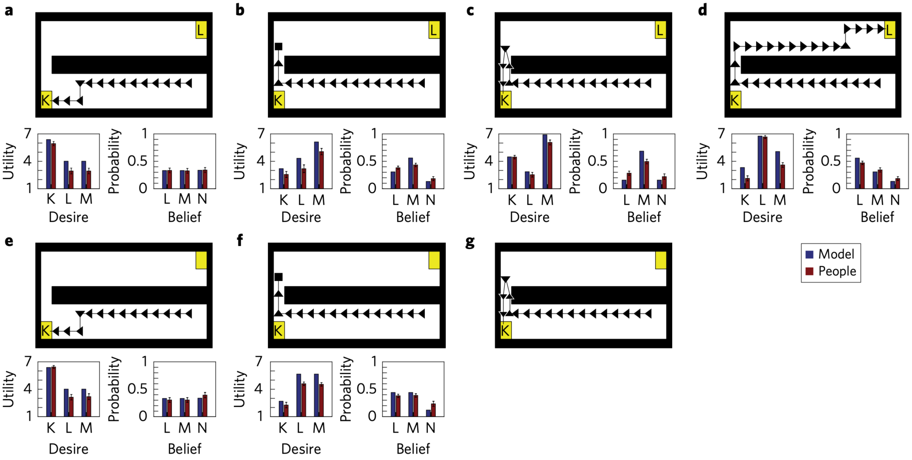
Experiment 1 Results (cont.)

Comparing BToM and mean human ($n = 16$) desire and belief inferences from seven key scenario types.



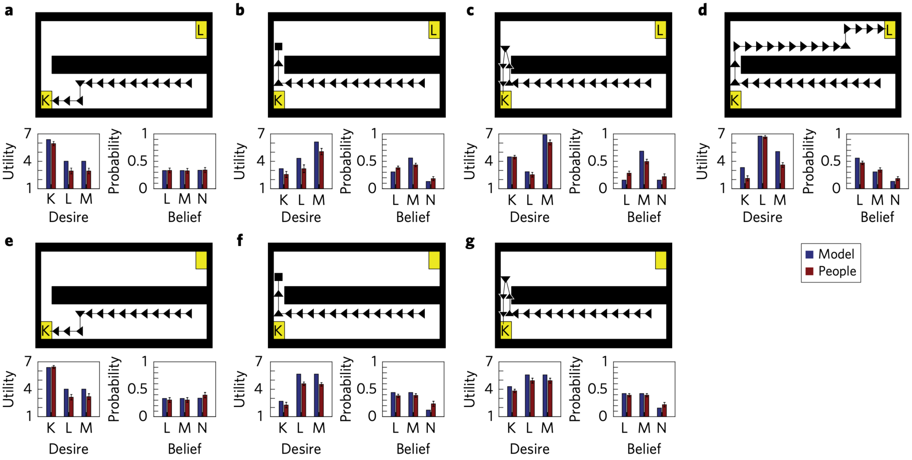
Experiment 1 Results (cont.)

Comparing BToM and mean human ($n = 16$) desire and belief inferences from seven key scenario types.



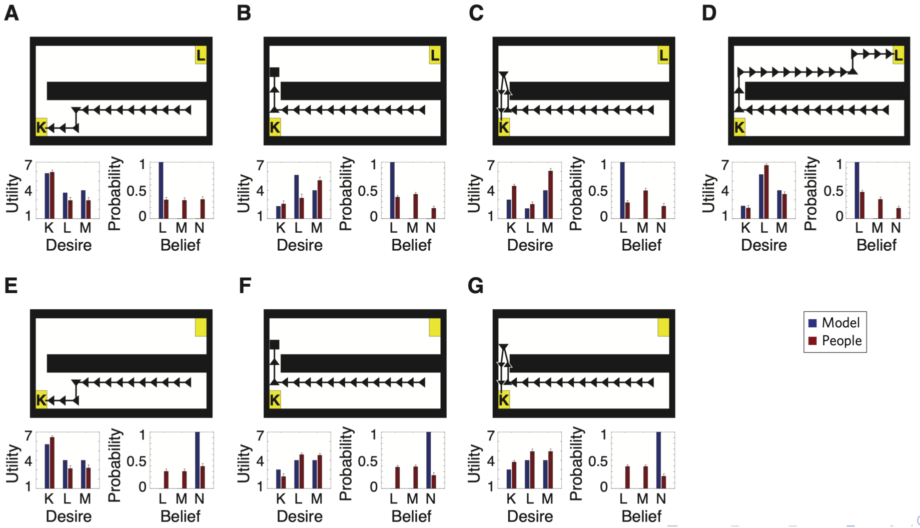
Experiment 1 Results (cont.)

Comparing BToM and mean human ($n = 16$) desire and belief inferences from seven key scenario types.



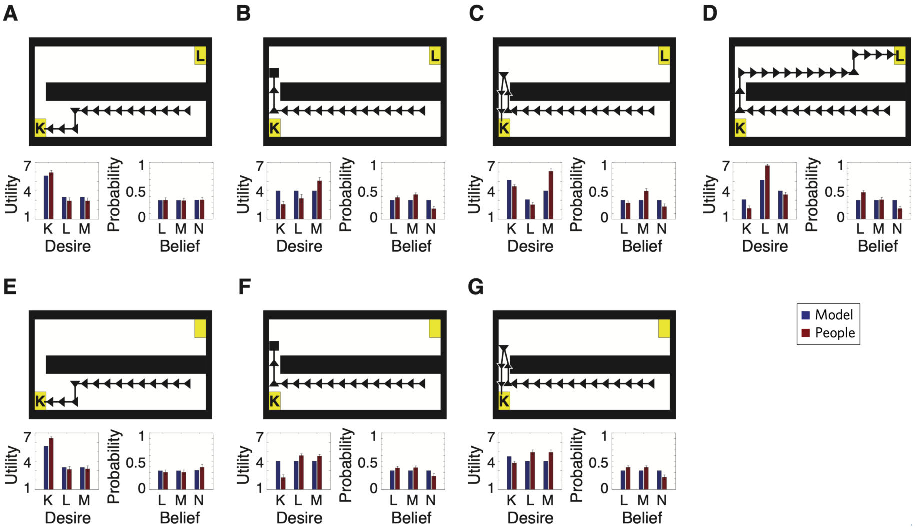
Experiment 1 Results (cont.)

Comparing TrueBelief and mean human ($n = 16$) desire and belief inferences from seven key scenario types.



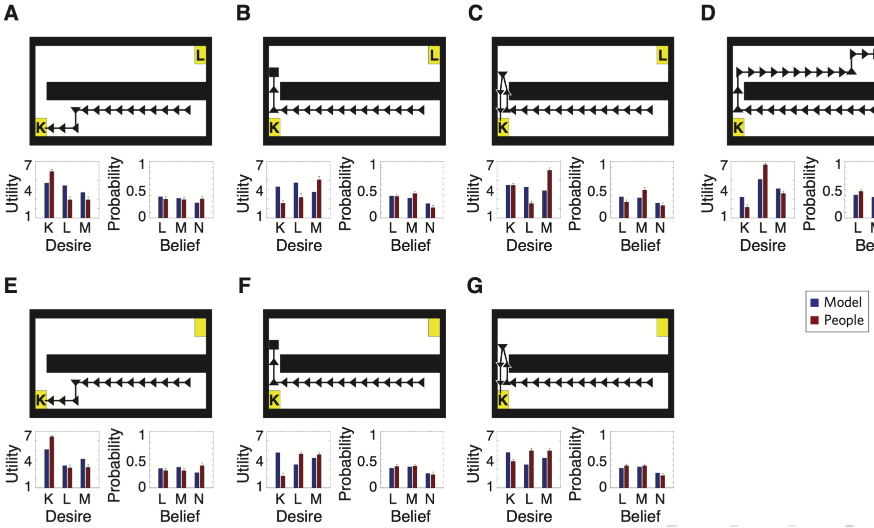
Experiment 1 Results (cont.)

Comparing NoCost and mean human ($n = 16$) desire and belief inferences from seven key scenario types.



Experiment 1 Results (cont.)

Comparing MotionHeuristic and mean human ($n = 16$) desire and belief inferences from seven key scenario types.



BToM vs. MotionHeuristic

Question 5

As a baseline, the BToM model is compared against a cue-based "MotionHeuristic" model, which only takes into account which objects the agent is moving towards / away from.

1. Why is the MotionHeuristic model unable to produce human-like inferences about the agent's *desires*?

BToM vs. MotionHeuristic

Question 5

As a baseline, the BToM model is compared against a cue-based "MotionHeuristic" model, which only takes into account which objects the agent is moving towards / away from.

1. Why is the MotionHeuristic model unable to produce human-like inferences about the agent's *desires*?
2. Why is MotionHeuristic better at producing human-like inferences about the agent's *beliefs*?

BToM vs. MotionHeuristic

Question 5

As a baseline, the BToM model is compared against a cue-based "MotionHeuristic" model, which only takes into account which objects the agent is moving towards / away from.

1. Why is the MotionHeuristic model unable to produce human-like inferences about the agent's *desires*?
2. Why is MotionHeuristic better at producing human-like inferences about the agent's *beliefs*?
3. How might the scenarios be modified to "break" the MotionHeuristic, so that it no longer produces human-like inferences about the agent's beliefs?

Experiment 2: Free Food Carts

- Food carts: Afghani (A), Burmese (B), and Colombian (C).

[1] Chris L. Baker et al., "Rational quantitative attribution of beliefs, desires and percepts in human mentalizing".

Experiment 2: Free Food Carts

- Food carts: Afghani (A), Burmese (B), and Colombian (C).
- Cart location: north (N), west (W), and east (E).

[1] Chris L. Baker et al., "Rational quantitative attribution of beliefs, desires and percepts in human mentalizing".

Experiment 2: Free Food Carts

- Food carts: Afghani (A), Burmese (B), and Colombian (C).
- Cart location: north (N), west (W), and east (E).
- Cart A and B can be open or closed. Cart C is always open.

[1] Chris L. Baker et al., "Rational quantitative attribution of beliefs, desires and percepts in human mentalizing".

Experiment 2: Free Food Carts

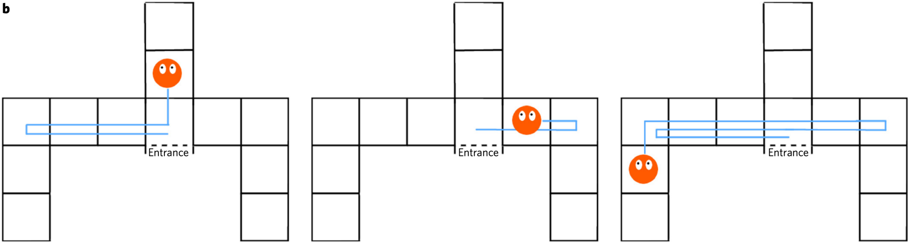
- Food carts: Afghani (A), Burmese (B), and Colombian (C).
- Cart location: north (N), west (W), and east (E).
- Cart A and B can be open or closed. Cart C is always open.
- The agent is a student finding free food with preference $A \succ B \succ C$.

[1] Chris L. Baker et al., "Rational quantitative attribution of beliefs, desires and percepts in human mentalizing".

Experiment 2: Free Food Carts

- Food carts: Afghani (A), Burmese (B), and Colombian (C).
- Cart location: north (N), west (W), and east (E).
- Cart A and B can be open or closed. Cart C is always open.
- The agent is a student finding free food with preference $A \succ B \succ C$.

The observers see the agent's path but not the cart locations or availabilities. They are required to infer the positions of all three carts.



[1] Chris L. Baker et al., "Rational quantitative attribution of beliefs, desires and percepts in human mentalizing".

BSCV used 100,000 iterations, with training folds containing 13/19 of scenario types, and testing folds containing 6/19 of scenario types. Values reported are median BSCV correlations with 95% confidence intervals. (*), (**) indicate r-values which are significantly less than those of BToM ($p < 0.0001$; $p < 0.001$).

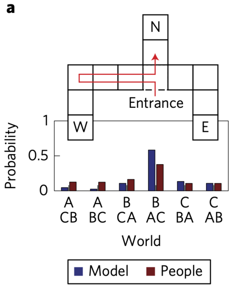
Table 3: BSCV analysis of model predictions for individual scenarios

r (BSCV)	BToM	TimeBelief	NoCost	MotionHeuristic
World State	**0.91 (0.86, 0.94)**	0.63 (0.24, 0.83)**	0.46 (0.17, 0.79)**	0.61 (0.10, 0.83)*

[1] Chris L. Baker et al., "Rational quantitative attribution of beliefs, desires and percepts in human mentalizing".

Experiment 2 Results (cont.)

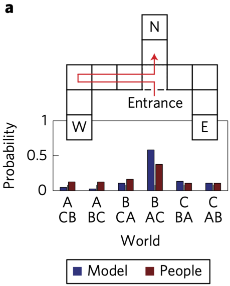
Comparing BToM and mean human ($n = 176$) percept inferences on a range of key scenarios.

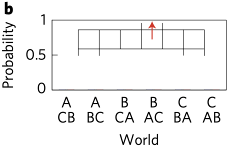


[1] Chris L. Baker et al., "Rational quantitative attribution of beliefs, desires and percepts in human mentalizing".

Experiment 2 Results (cont.)

Comparing BToM and mean human ($n = 176$) percept inferences on a range of key scenarios.

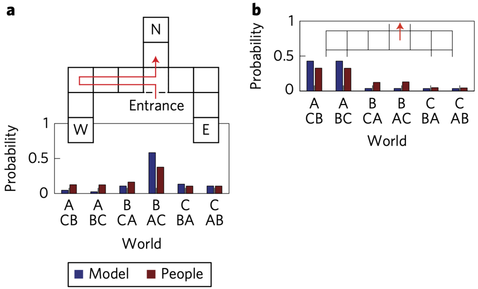




[1] Chris L. Baker et al., "Rational quantitative attribution of beliefs, desires and percepts in human mentalizing".

Experiment 2 Results (cont.)

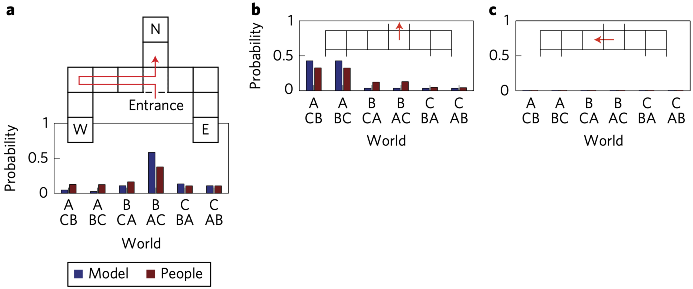
Comparing BToM and mean human ($n = 176$) percept inferences on a range of key scenarios.



[1] Chris L. Baker et al., "Rational quantitative attribution of beliefs, desires and percepts in human mentalizing".

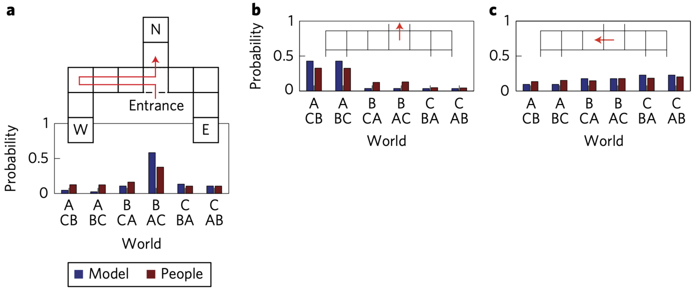
Experiment 2 Results (cont.)

Comparing BToM and mean human ($n = 176$) percept inferences on a range of key scenarios.



[1] Chris L. Baker et al., "Rational quantitative attribution of beliefs, desires and percepts in human mentalizing".

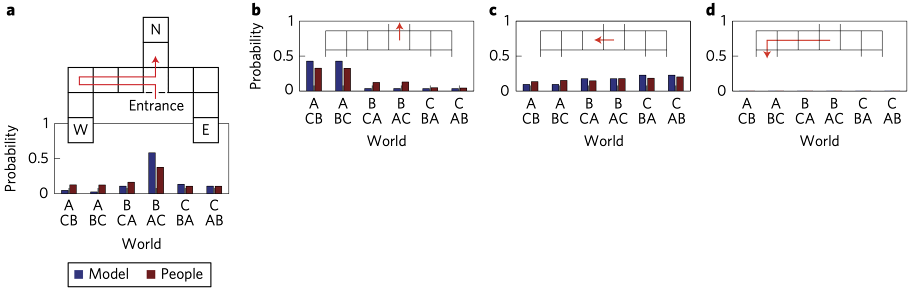
Comparing BToM and mean human ($n = 176$) percept inferences on a range of key scenarios.



[1]Chris L. Baker et al., "Rational quantitative attribution of beliefs, desires and percepts in human mentalizing".

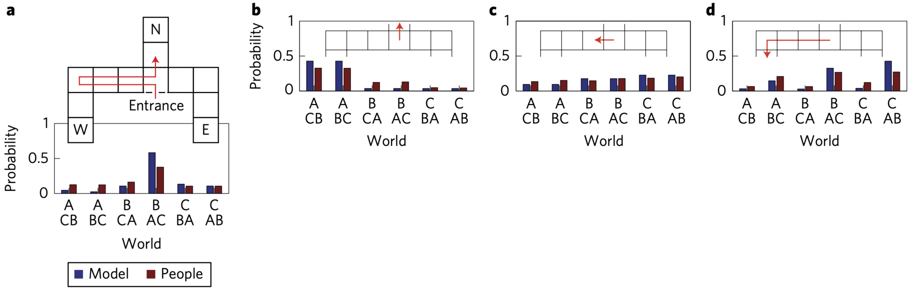
Experiment 2 Results (cont.)

Comparing BToM and mean human ($n = 176$) percept inferences on a range of key scenarios.



[1] Chris L. Baker et al., "Rational quantitative attribution of beliefs, desires and percepts in human mentalizing".

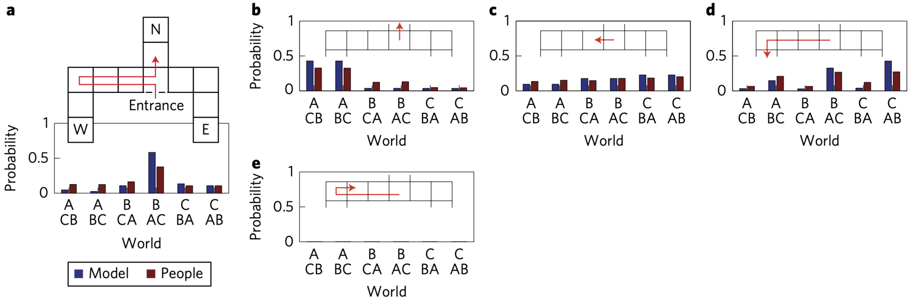
Comparing BToM and mean human ($n = 176$) percept inferences on a range of key scenarios.



[1] Chris L. Baker et al., "Rational quantitative attribution of beliefs, desires and percepts in human mentalizing".

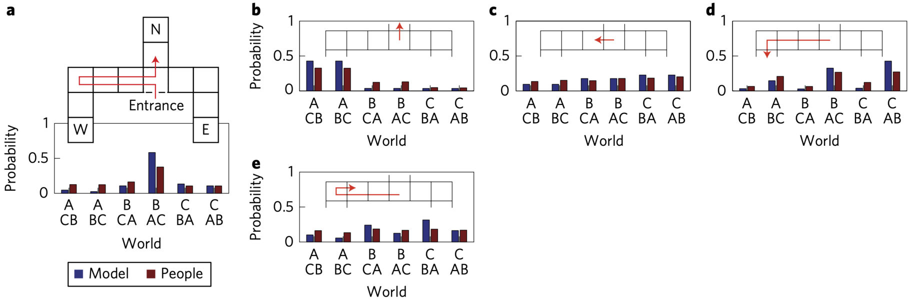
Experiment 2 Results (cont.)

Comparing BToM and mean human ($n = 176$) percept inferences on a range of key scenarios.



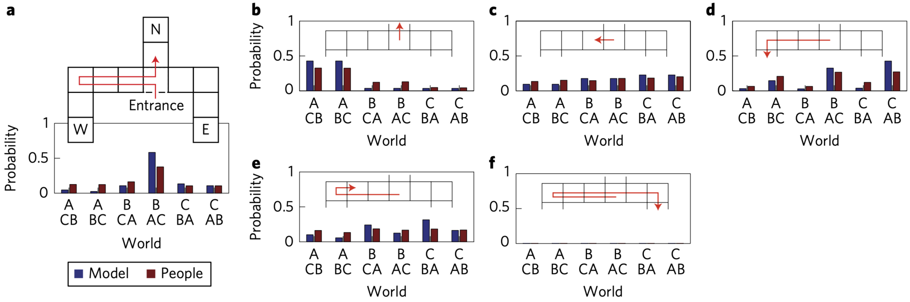
[1] Chris L. Baker et al., "Rational quantitative attribution of beliefs, desires and percepts in human mentalizing".

Comparing BToM and mean human ($n = 176$) percept inferences on a range of key scenarios.



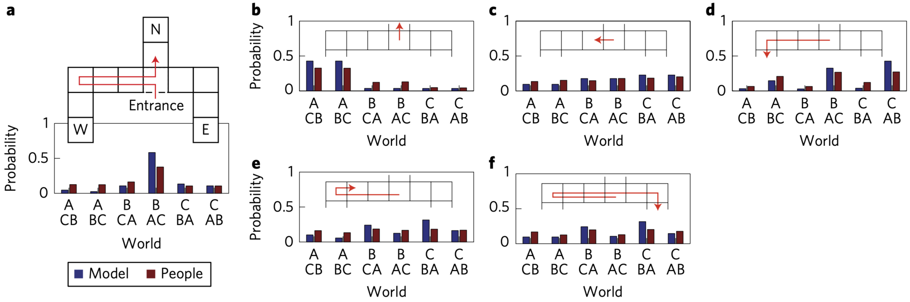
[1] Chris L. Baker et al., "Rational quantitative attribution of beliefs, desires and percepts in human mentalizing".

Comparing BToM and mean human ($n = 176$) percept inferences on a range of key scenarios.



[1] Chris L. Baker et al., "Rational quantitative attribution of beliefs, desires and percepts in human mentalizing".

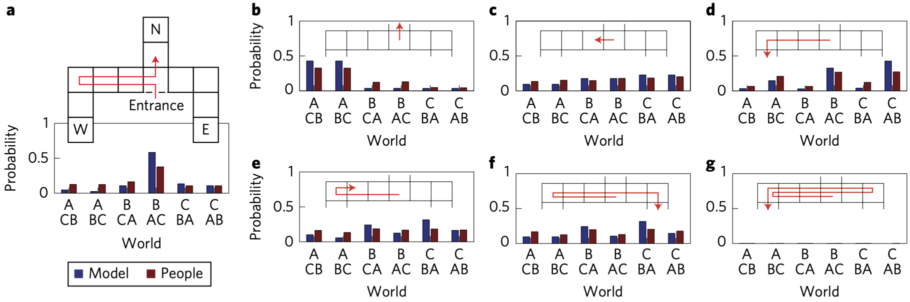
Comparing BToM and mean human ($n = 176$) percept inferences on a range of key scenarios.



[1] Chris L. Baker et al., "Rational quantitative attribution of beliefs, desires and percepts in human mentalizing".

Experiment 2 Results (cont.)

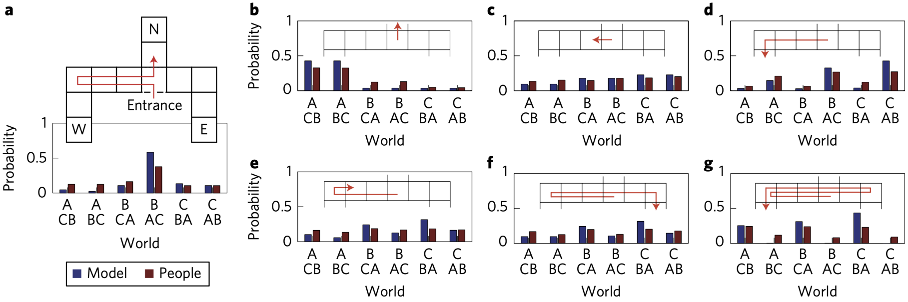
Comparing BToM and mean human ($n = 176$) percept inferences on a range of key scenarios.



[1] Chris L. Baker et al., "Rational quantitative attribution of beliefs, desires and percepts in human mentalizing".

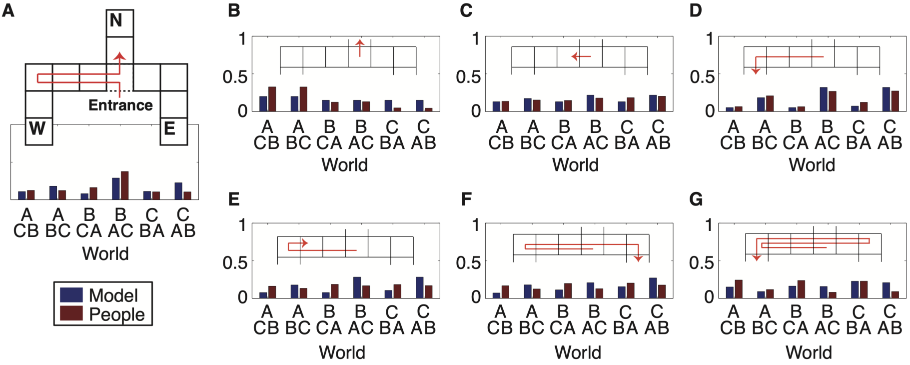
Experiment 2 Results (cont.)

Comparing BToM and mean human ($n = 176$) percept inferences on a range of key scenarios.



[1] Chris L. Baker et al., “Rational quantitative attribution of beliefs, desires and percepts in human mentalizing”.

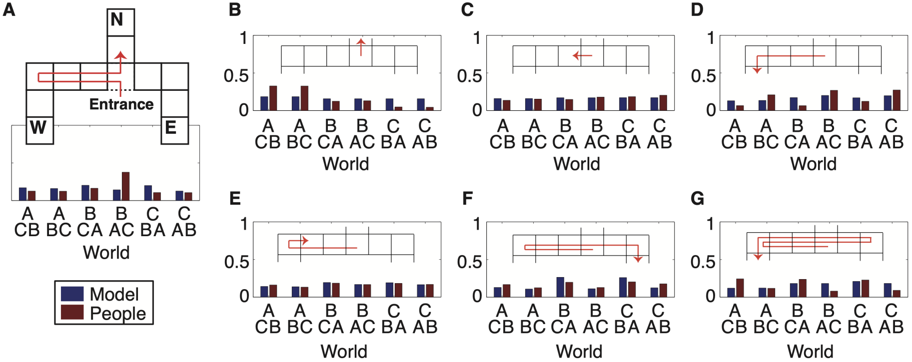
Comparing TrueBelief and mean human ($n = 176$) percept inferences on a range of key scenarios.



[1] Chris L. Baker et al., "Rational quantitative attribution of beliefs, desires and percepts in human mentalizing".

Experiment 2 Results (cont.)

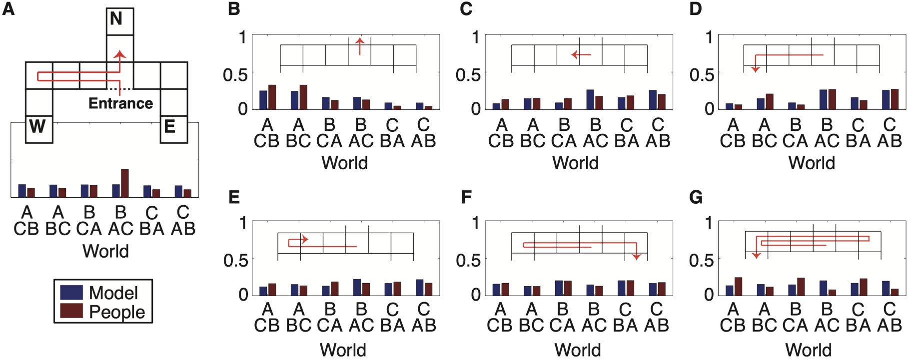
Comparing NoCost and mean human ($n = 176$) percept inferences on a range of key scenarios.



[1] Chris L. Baker et al., "Rational quantitative attribution of beliefs, desires and percepts in human mentalizing".

Experiment 2 Results (cont.)

Comparing MotionHeuristic and mean human ($n = 176$) percept inferences on a range of key scenarios.



[1] Chris L. Baker et al., "Rational quantitative attribution of beliefs, desires and percepts in human mentalizing".

Broader Insights

Question 6

To the extent that LLMs can replicate theory-of-mind-associated capabilities like attributing beliefs and desires, do you think they are model-based or cue-based (or something in between)? How could we design experiments to tell?

Broader Insights

Question 6

To the extent that LLMs can replicate theory-of-mind-associated capabilities like attributing beliefs and desires, do you think they are model-based or cue-based (or something in between)? How could we design experiments to tell?

Question 7: Bonus

What is the POMDP solver used for planning in Experiment 2, and how is it related to NUS?

Table of Contents

Summary

- **Bayesian Theory of Mind (BToM)** is a model for inferring others' beliefs, desires, and percepts.
- Quantitatively evaluated in two experiments:
  - **Experiment 1:** Predicting human inferences about beliefs and desires from action trajectories.
  - **Experiment 2:** Inferring hidden aspects of the world (percepts) from others' actions.
- BToM outperforms alternative models (TrueBelief, NoCost, MotionHeuristic), capturing both quantitative fits and qualitative nuances in human judgments.

Future Developments

- Extend BToM to handle:
  - **Epistemic goals** (explicit information-seeking) in addition to instrumental goals.
  - **Multi-agent interactions** (competitive/cooperative scenarios) using game-theoretic models.
  - **Richer environment models** with intuitive physics and broader action repertoires.
- Integrate **fast, learned approximations** (e.g., neural networks) for real-time inference.

- THE END -

Thank you for your attention

Acknowledgements

Special thanks to Prof. Tan Zhi Xuan for guidance and support throughout this presentation.

Contact

nqduc@u.nus.edu