

# UC Merced

## Proceedings of the Annual Meeting of the Cognitive Science Society

### Title

Modeling the Mistakes of Boundedly Rational Agents Within a Bayesian Theory of Mind

### Permalink

<https://escholarship.org/uc/item/7tr2w3c9>

### Journal

Proceedings of the Annual Meeting of the Cognitive Science Society, 43(43)

### Authors

Alanqary, Alwa

Lin, Gloria Z

Le, Joie

et al.

### Publication Date

2021

### Copyright Information

This work is made available under the terms of a Creative Commons Attribution License, available at <https://creativecommons.org/licenses/by/4.0/>

Peer reviewed

# Modeling the Mistakes of Boundedly Rational Agents Within a Bayesian Theory of Mind

**Alwa Alanqary**

Massachusetts Institute of Technology, Cambridge, Massachusetts, United States

**Gloria Lin**

Massachusetts Institute of Technology, Cambridge, Massachusetts, United States

**Joie Le**

Massachusetts Institute of Technology, Cambridge, Massachusetts, United States

**Tan Zhi-Xuan**

Massachusetts Institute of Technology, Cambridge, Massachusetts, United States

**Vikash Mansinghka**

MIT, Cambridge, Massachusetts, United States

**Josh Tenenbaum**

MIT, Cambridge, Massachusetts, United States

## Abstract

When inferring the goals that others are trying to achieve, people intuitively understand that others might make mistakes along the way. This is crucial for activities such as teaching, offering assistance, and deciding between blame or forgiveness. However, Bayesian models of theory of mind have generally not accounted for these mistakes, instead modeling agents as mostly optimal in achieving their goals. As a result, they are unable to explain phenomena like locking oneself out of one's house, or losing a game of chess. Here, we extend the Bayesian Theory of Mind framework to model boundedly rational agents who may have mistaken goals, plans, and actions. We formalize this by modeling agents as probabilistic programs, where goals may be confused with semantically similar states, plans may be misguided due to resource-bounded planning, and actions may be unintended due to execution errors. We present experiments eliciting human goal inferences in two domains: (i) a gridworld puzzle with gems locked behind doors, and (ii) a block-stacking domain. Our model better explains human inferences than alternatives, while generalizing across domains. These findings indicate the importance of modeling others as bounded agents, in order to account for the full richness of human intuitive psychology.