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

Arwa Alanqary, Gloria Z. Lin, Joie Le, Tan Zhi-Xuan, Vikash K. Mansinghka, Joshua B. Tenenbaum

Massachusetts Institute of Technology

Introduction

- Humans intuitively understand that others are fallible and might make mistakes.
- This allows them to infer the goals of others even from mistaken or failed plans.
- What explains this ability, and how can it be modeled?

Hypothesis: *Humans intuitively understand that others make mistakes because we are, at best, boundedly rational.*

Computational Model

- We specify a boundedly rational agent model using a probabilistic program. The model accounts for mistaken goals, plans, and actions.

```
model agent_step(t, goal, plan, state):

Mistaken Goals
via temporary goal
confusion          { Corrupt or repair goal with probability ε_g.

                    if goal is reached:
                      return nothing
Mistaken Plans      else if length(plan) < t or state is unexpected:
via resource-         Sample search budget for planner.
bounded planning    { Search for partial plan from current state to goal.
                      Extend plan with partial plan.

Mistaken Actions    with probability 1-ε_a:
via execution errors { Select action from timestep t of plan.
                    else:
                      Randomly choose an action available in state.

                    return action
```

- We model human observers as performing Bayesian goal inference over this model, given a series of observations of the agent and its environment
- Inference done via Sequential Inverse Plan Search, a particle filtering algorithm. (Zhi-Xuan et al, NeurIPS 2020)

Experiments

- We elicited goal inferences from human participants as they watched a variety of optimal and suboptimal agent trajectories unfold.
- Trajectories were designed to exhibit mistakes in the agent's *goals*, *plans* or *actions*. 16 trajectories per domain, 4 optimal, 4 for each mistake type.
- 32 participants for Doors Keys & Gems, 20 participants for Block Words, recruited from MTurk.

Domains

- To demonstrate the generality of our model, we conducted experiments in two domains:
  1. A gridworld puzzle called **Doors, Keys & Gems**
  2. A Blocks World variant called **Block Words** where an agent spells words out of lettered blocks

.

- These domains exhibit the compositional structure that humans encounter in daily life, making them tractable to plan in, but also complex enough for mistakes to arise.

Baselines

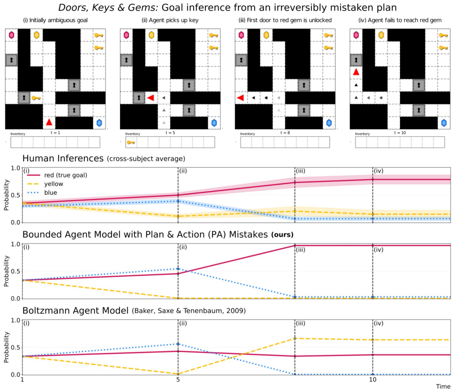
G-lesioned: mistaken **goals** are not modeled
Agents always plan to achieve their original intended goals

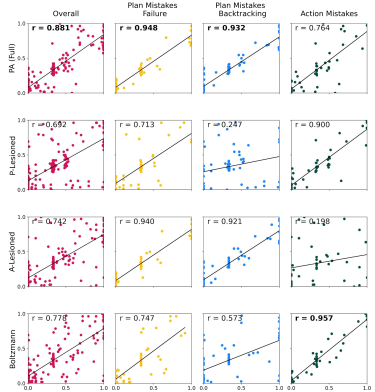
P-lesioned: mistaken **plans** are not modeled
Agents start off with an optimal plan to the goal, and form new optimal plans after making action mistakes

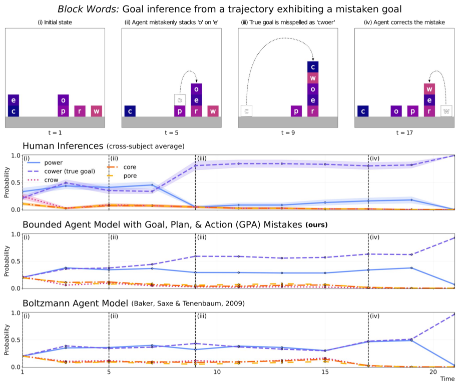
A-lesioned: mistaken **actions** are not modeled
Agents always execute actions according to their plans.

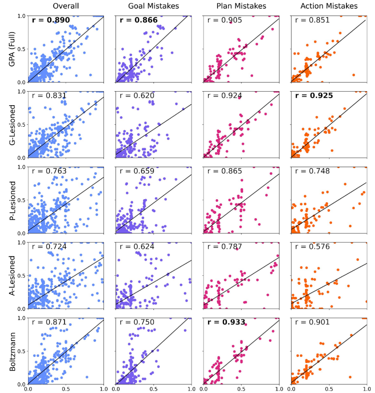
Boltzmann agent model (Baker, Saxe & Tenenbaum 2009)
Agents compute expected future reward of every state
Actions are selected according to a Boltzmann distribution

Results


Doors, Keys & Gems: Goal inference from an irreversibly mistaken plan




Block Words: Goal inference from a trajectory exhibiting a mistaken goal


Overall
Goal Mistakes
Plan Mistakes
Action Mistakes
GPA (Full)
G-Lesioned
P-Lesioned
A-Lesioned
Boltzmann
r = 0.890
r = 0.866
r = 0.905
r = 0.851
r = 0.831
r = 0.620
r = 0.924
r = 0.925
r = 0.763
r = 0.659
r = 0.865
r = 0.748
r = 0.724
r = 0.624
r = 0.787
r = 0.576
r = 0.871
r = 0.750
r = 0.933
r = 0.901