[1] NYU

[2] TOYOTA RESEARCH INSTITUTE

Estimating Cognitive Biases with Attention-Aware Inverse Planning

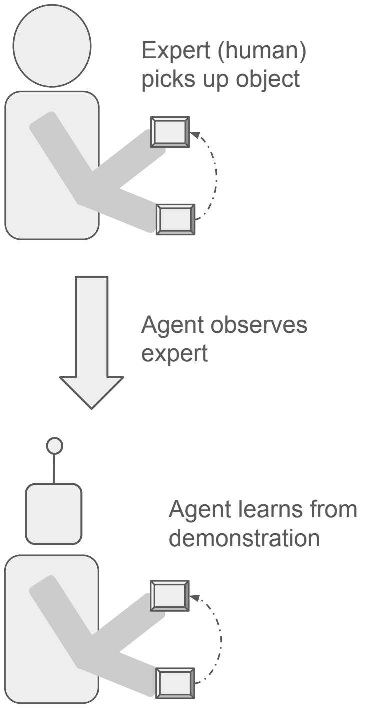
Sounak Banerjee[1] Daphne Cornelisse[1] Deepak Gopinath[2]
Emily Sumner[2] Jonathan DeCastro[2] Guy Rosman[2] Eugene Vinitsky[1]
Mark K. Ho[1]

Computation and Decision-Making Lab

EMERGE

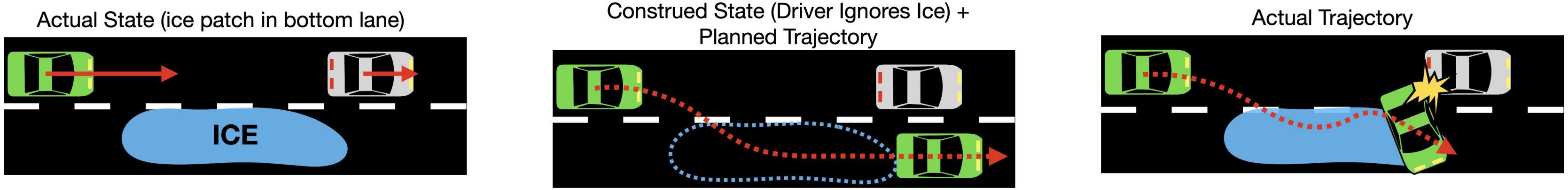
Motivation: Challenges for Models of Human Behavior

- Inverse Reinforcement Learning (IRL): Infer reward structures from expert behavior, then compute an optimal policy using the inferred rewards.
(Ng and Russell, 2000)
    - Limitation: These approaches struggle when there is a mismatch between the expert's and learner's transition dynamics (Viano et al., 2020)



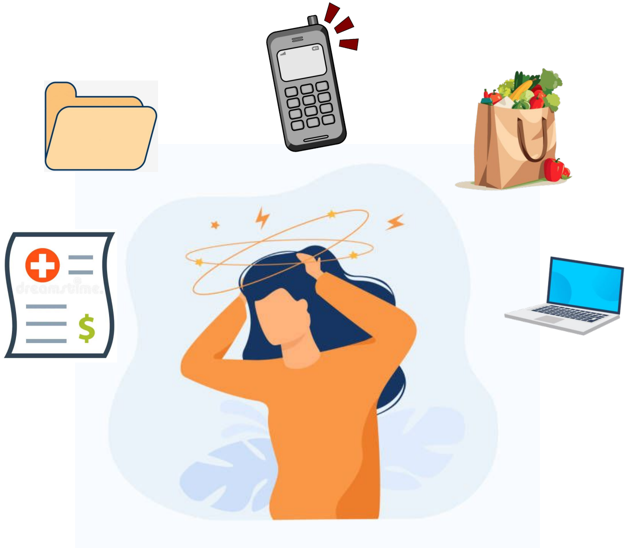
Motivation: Challenges for Models of Human Behavior

- Cognitive Limitations: Agents with limited cognitive capacities rely on simplified world models (different transition dynamics) to compute action plans.



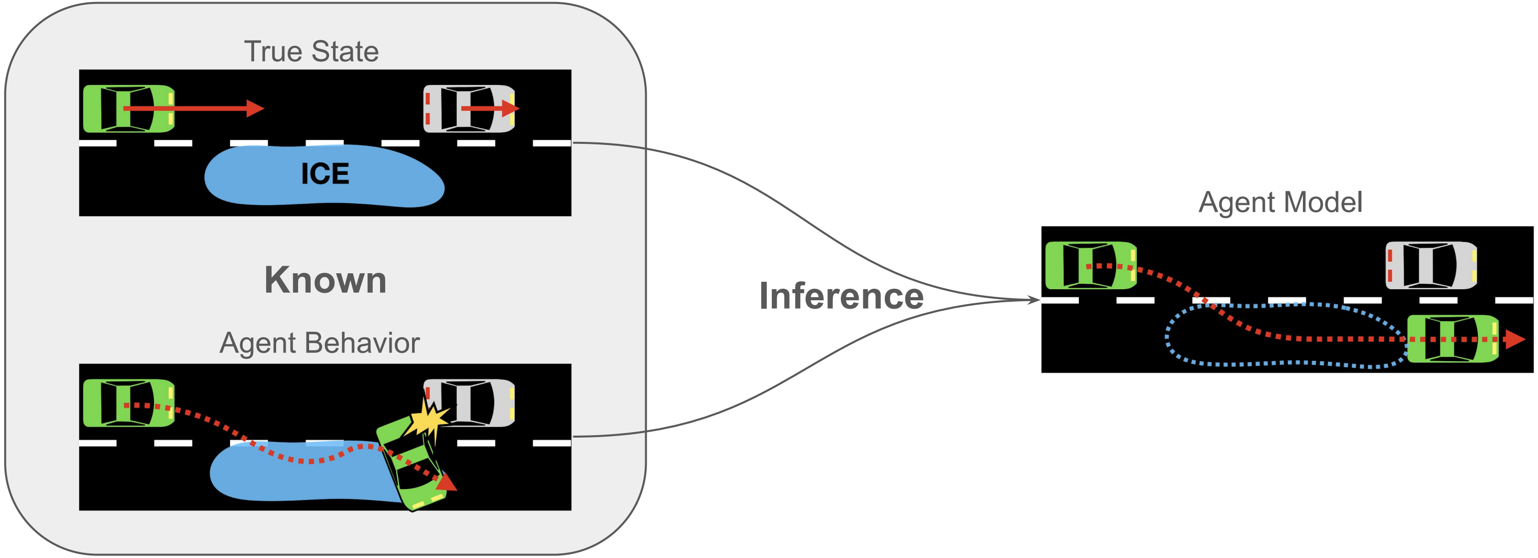
Motivation: Challenges for Models of Human Behavior

- Humans are **boundedly rational** agents (Simon, 1957; Griffiths et al., 2015; Gershman et al., 2015)
    - Limited resources such as cognitive capacity, which include attention capabilities
- Models of human behavior should be able to **capture systematic suboptimalities** (such as limited attention) from behavior (Ho et al., 2022)



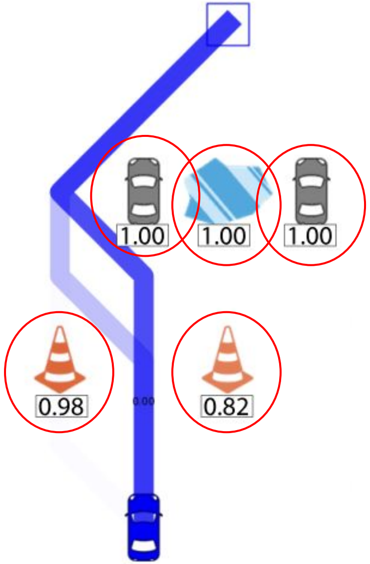
Attention Aware Inverse Planning (AAIP)

- An approach for estimating attentional heuristics of attention-limited agents, from observed behavior
  - Inverting the planning process of an attention-limited agent

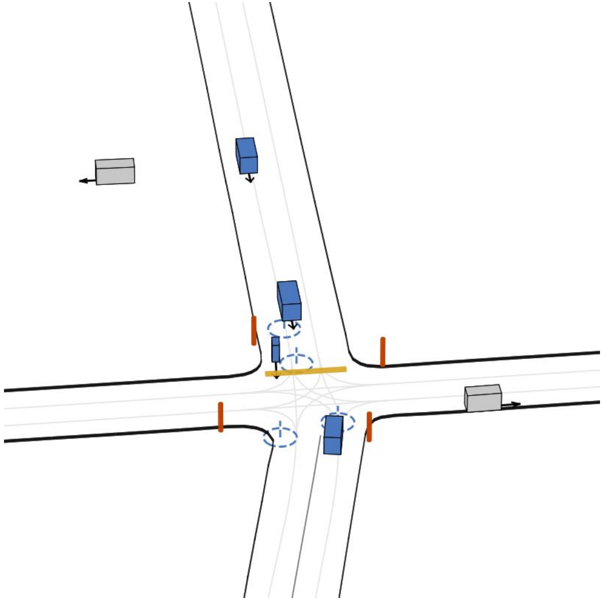


Simulators Used

Properties:
- Simple **MDP-based discrete** environment dynamics
- Optimal policy obtained using **value iteration**



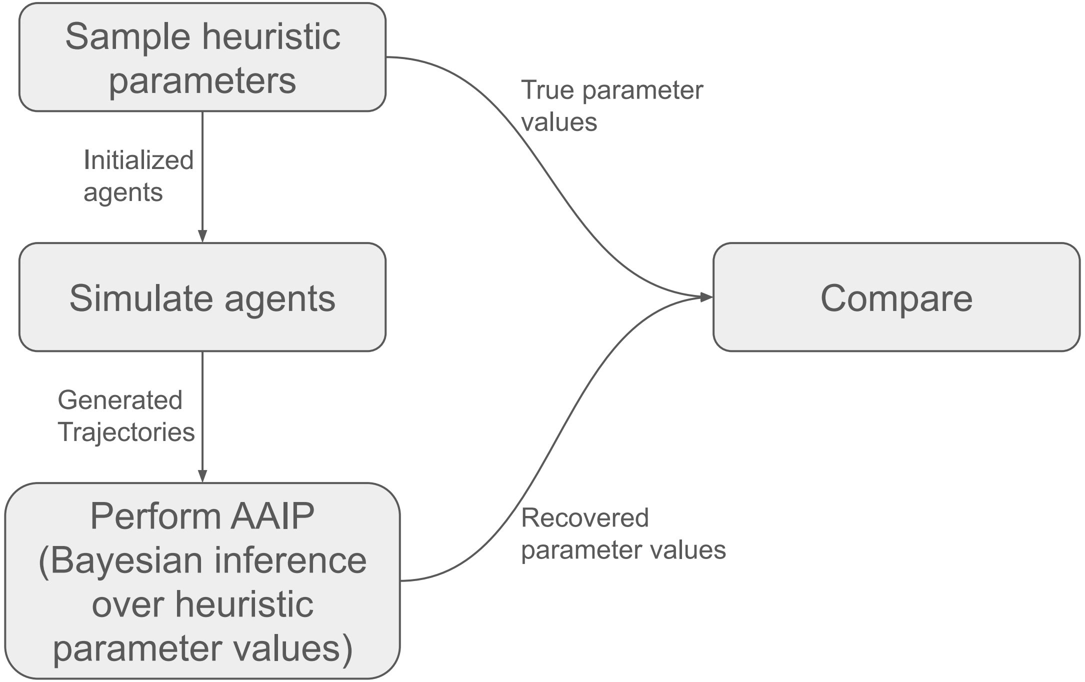
Driving World



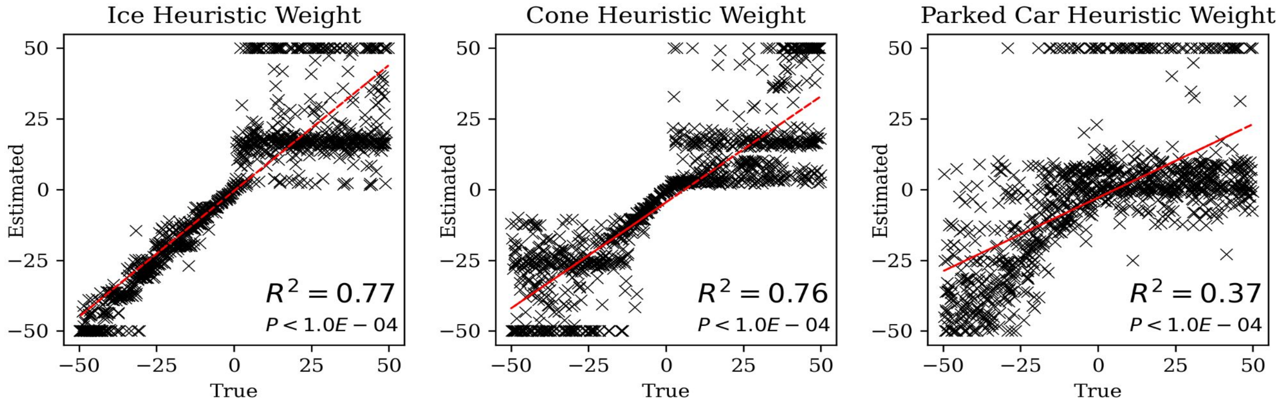
GPUDrive

Properties:
- **Complex and continuous** environment dynamics
- Optimal policy obtained using **deep reinforcement learning**

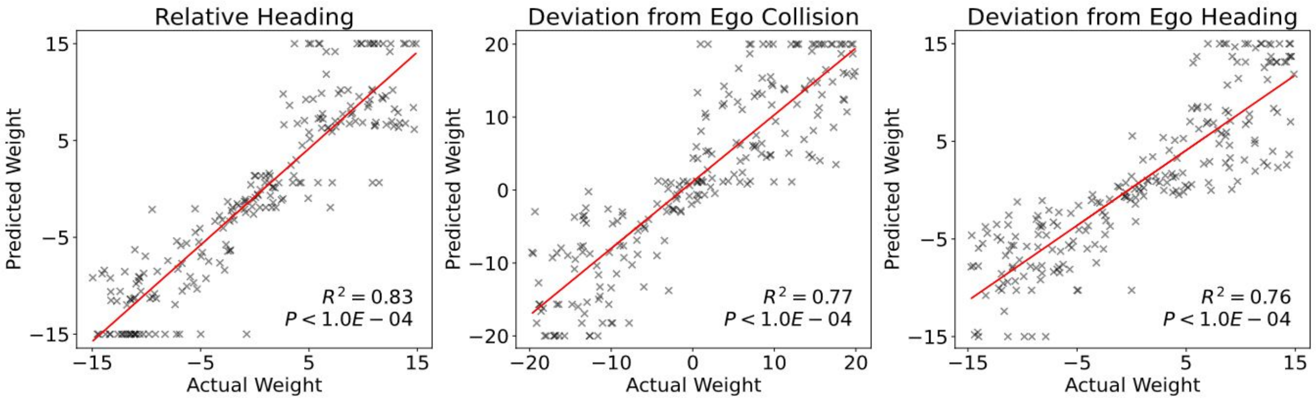
Experiment Pipeline



Results



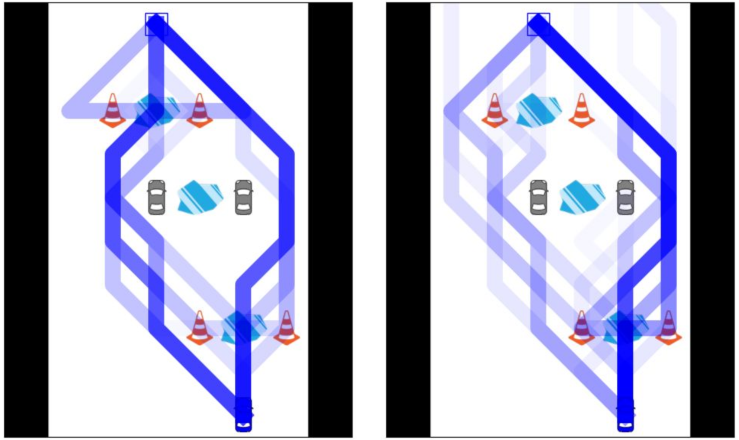
Driving World



GPUDrive

Results: AAIP vs IRL in DrivingWorld

Attention-limited decision-maker policy | Inferred IRL policy



	β	ε	$\tilde{r}_{\text{Ice}}$	$\tilde{r}_{\text{Ice+Cone}}$	γ	NLL↓
Noise	×	×				555
IRL	×	×	×			537
AAIP	×	×	NA	NA	NA	**265**

Summary

- Formally introduce the Attention-Aware Inverse Planning problem
- Demonstrate how AAIP systematically differs from standard IRL
- Leverage deep reinforcement learning with computational cognitive modeling to AAIP problems
- Show that our approach can reliably and accurately capture behavior of attention-limited agents

Code

CoDec Lab

TRI