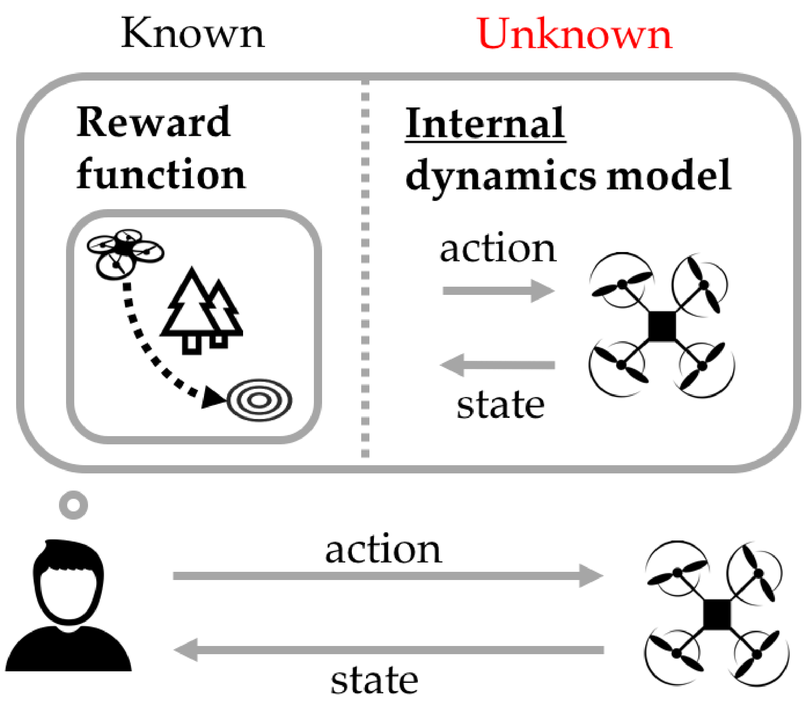
# Where Do You Think You're Going?: Inferring Beliefs about Dynamics from Behavior

Siddharth Reddy, Anca D. Dragan, Sergey Levine

{sgr,anca,svlevine}@berkeley.edu

BERKELEY ARTIFICIAL INTELLIGENCE RESEARCH

## Explaining Demonstrated Actions




## Learning the Internal Dynamics Model

**Demonstrations** → (Our learning algorithm) → **Demonstrator's internal dynamics model**

In an MDP with a discrete action space $\mathcal{A}$, the human demonstrator is assumed to follow a policy $\pi$ that maximizes an entropy-regularized reward $R(s, a, s')$ under dynamics $T(s'|s, a)$. Equivalently,

$$\pi(a|s) \triangleq \frac{\exp\left(Q(s,a)\right)}{\sum_{a' \in \mathcal{A}} \exp\left(Q(s,a')\right)}, \tag{1}$$

where $Q$ is the soft $Q$ function, which satisfies the soft Bellman equation,

$$Q(s,a) = \mathbb{E}_{s' \sim T(\cdot|s,a)}\left[R(s,a,s') + \gamma V(s')\right], \tag{2}$$

with $V$ the soft value function,

$$V(s) \triangleq \log\left(\sum_{a \in \mathcal{A}} \exp\left(Q(s,a)\right)\right). \tag{3}$$

Soft Bellman error:

$$\delta_i(s,a) \triangleq Q_i(s,a) - \int_{s' \in \mathcal{S}} T(s'|s,a)\left(R_i(s,a,s') + \gamma V_i(s')\right) \mathrm{d}s'. \tag{4}$$

**Constrained optimization problem**:

$$\underset{\{\boldsymbol{\theta_i}\}_{i=1}^n, \boldsymbol{\phi}}{\text{minimize}} \sum_{i=1}^{n} \sum_{(s,a) \in \mathcal{D}_i^{\text{demo}}} -\log \pi_{\boldsymbol{\theta_i}}(a|s) \tag{5}$$
$$\text{subject to} \quad \delta_{\boldsymbol{\theta_i}, \boldsymbol{\phi}}(s,a) = 0 \ \forall i \in \{1, 2, ..., n\}, s \in \mathcal{S}, a \in \mathcal{A}.$$

Loss function for unconstrained optimization problem:

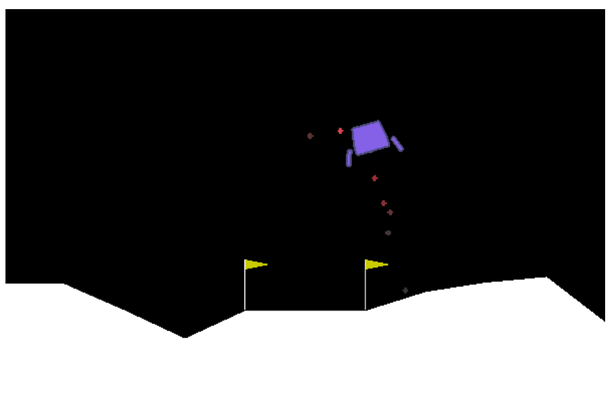
$$c(\boldsymbol{\theta}, \boldsymbol{\phi}) \triangleq \sum_{i=1}^{n} \sum_{(s,a) \in \mathcal{D}_i^{\text{demo}}} -\log \pi_{\boldsymbol{\theta_i}}(a|s) + \frac{\rho}{2} \sum_{i=1}^{n} \int_{s \in \mathcal{S}} \sum_{a \in \mathcal{A}} (\delta_{\boldsymbol{\theta_i}, \boldsymbol{\phi}}(s,a))^2 \mathrm{d}s. \tag{6}$$

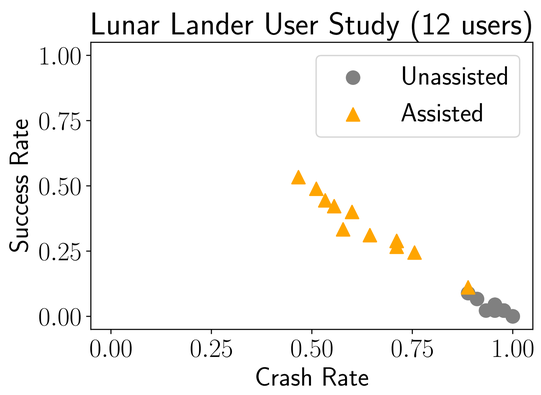
**Regularization**: multiple training tasks, and the action intent prior,

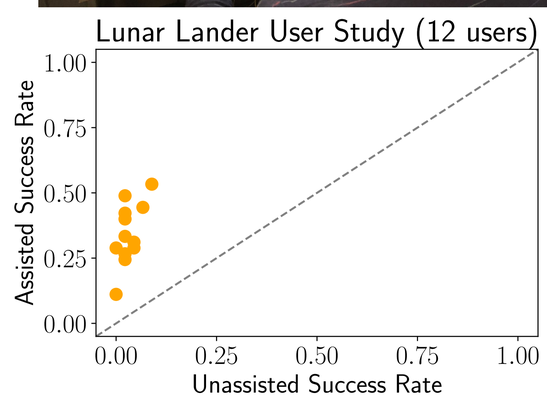
$$T_{\boldsymbol{\phi}}(s'|s,a) \triangleq \sum_{a^{\text{int}} \in \mathcal{A}} T^{\text{real}}(s'|s,a^{\text{int}}) f_{\boldsymbol{\phi}}(a^{\text{int}}|s,a). \tag{7}$$

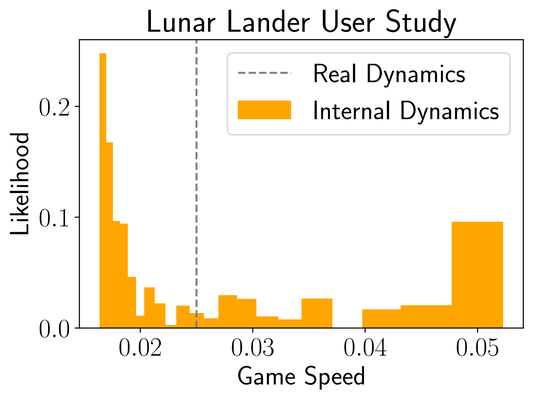
## User Study

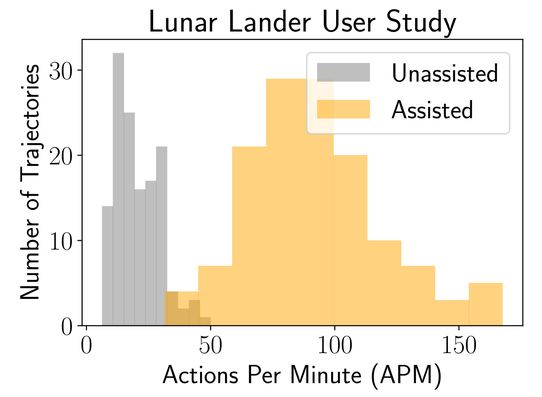
We asked 12 participants to play the **Lunar Lander game** without and with **internal-to-real dynamics transfer** assistance.

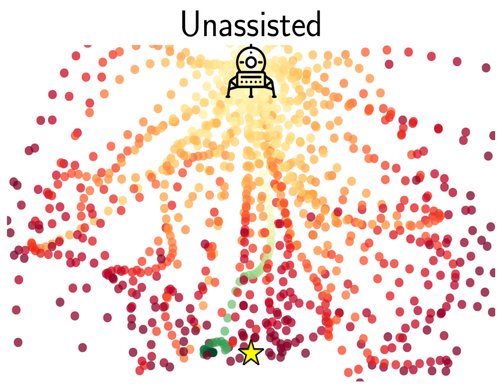


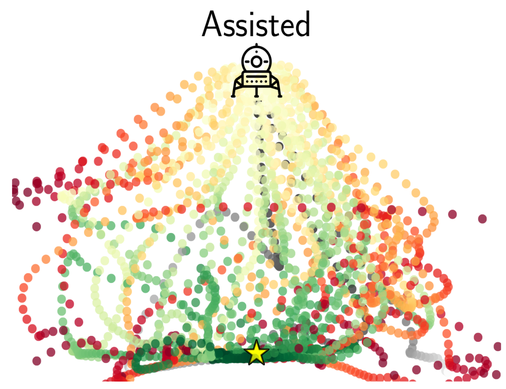


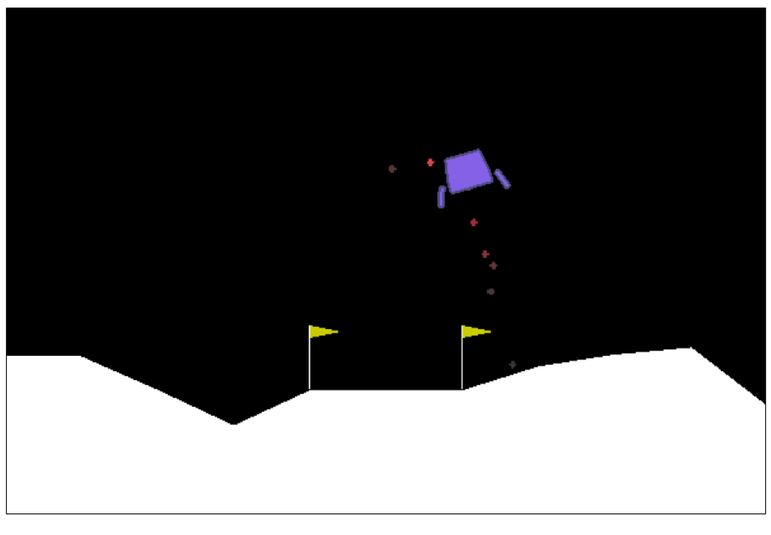



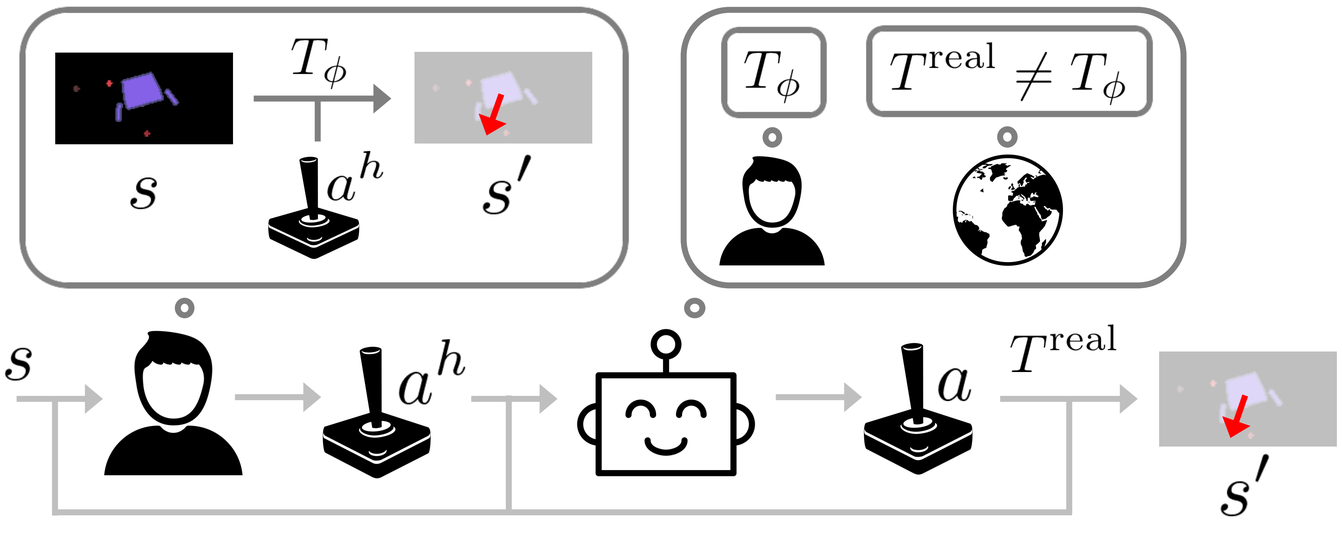




Means reported below for responses on a 7-point Likert scale, where 1 = Strongly Disagree, 4 = Neither Disagree nor Agree, and 7 = Strongly Agree.

| | $p$-value | Unassisted | Assisted |
|---|---|---|---|
| I enjoyed playing the game | $< .001$ | 3.92 | **5.92** |
| I improved over time | $< .0001$ | 3.08 | **5.83** |
| I didn't crash | $< .001$ | 1.17 | **3.00** |
| I didn't fly out of bounds | $< .05$ | 1.67 | **3.08** |
| I didn't run out of time | $> .05$ | 5.17 | 6.17 |
| I landed between the flags | $< .001$ | 1.92 | **4.00** |
| I understood how to complete the task | $< .05$ | 6.42 | **6.75** |
| I intuitively understood the physics of the game | $< .01$ | 4.58 | **6.00** |
| My actions were carried out | $> .05$ | 4.83 | 5.50 |
| My intended actions were carried out | $< .01$ | 2.75 | **5.25** |






Videos, code, and data available at `https://sites.google.com/view/inferring-internal-dynamics`