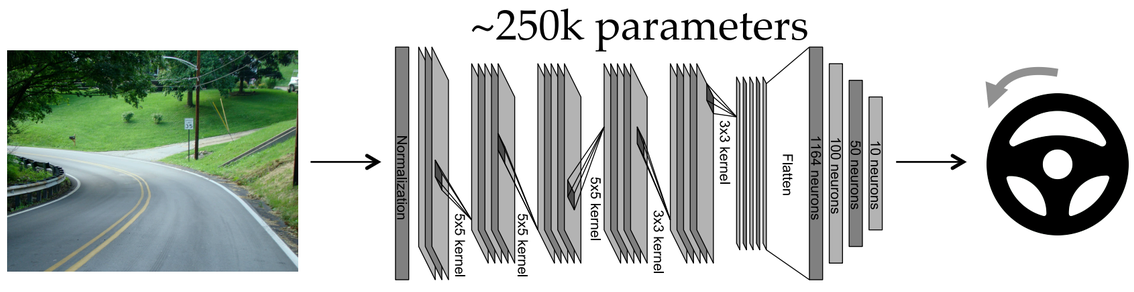
Leveraging Critical States to Develop Trust

Sandy Huang[1], Kush Bhatia[1], Pieter Abbeel[1,2], Anca Dragan[1]

[1]UC Berkeley, [2]OpenAI

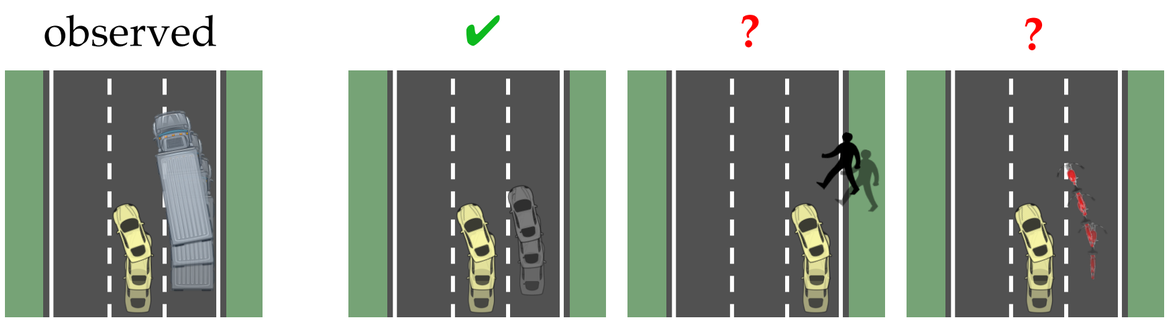
Barriers to Developing Trust

Complex tasks require complex policies



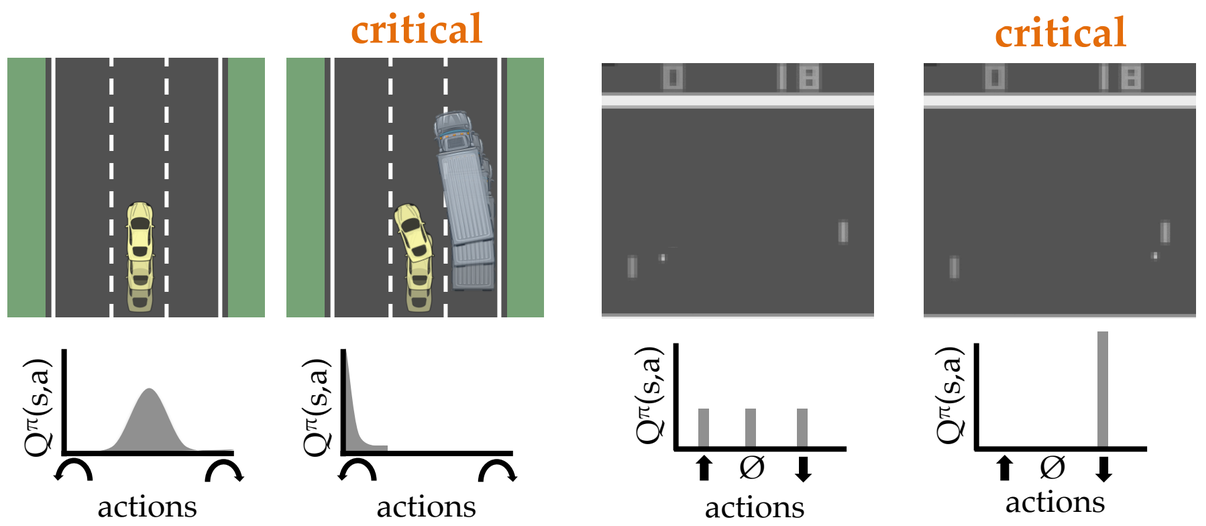
Bojarski et al., arXiv 2016

but users may be reluctant to trust such policies, due to limited extrapolation



Critical States Matter More

Hypothesis: If the robot shows us how it acts in critical states, we would gain a better appreciation of the policy, and if we agree with these actions, we would be more willing to trust the robot.



Algorithmic Teaching for Trust

Definition of trust:

$$\frac{1}{|S|}\sum_{s\in S} V^{\hat{\pi}_{E_i}}(s) = \frac{1}{|S|}\sum_{s\in S}\sum_{a\in A} \hat{\pi}_{E_i}(a|s)\, Q^{\hat{\pi}_{E_i}}(s,a)$$

average cumulative discounted reward obtained by human's policy estimate

Human model:

$$\hat{\pi}_{E_n}(a|s) = \begin{cases} g_{s^*,s}(\pi(s^*)) & \text{if } d(s,s^*) \le \beta, \\ & s^* = \arg\min_{s_i \in E_n} d(s, s_i) \\ \text{unif}(\mathcal{A}) & \text{otherwise.} \end{cases}$$

given $(s, \pi(s))$, human infers this is also the action distribution in all states similar to s

$$d(s,s') = \begin{cases} 0 & \text{if states are in the same cluster} \\ 1 & \text{otherwise} \end{cases}$$

human initially believes robot acts *randomly*

Selecting examples:

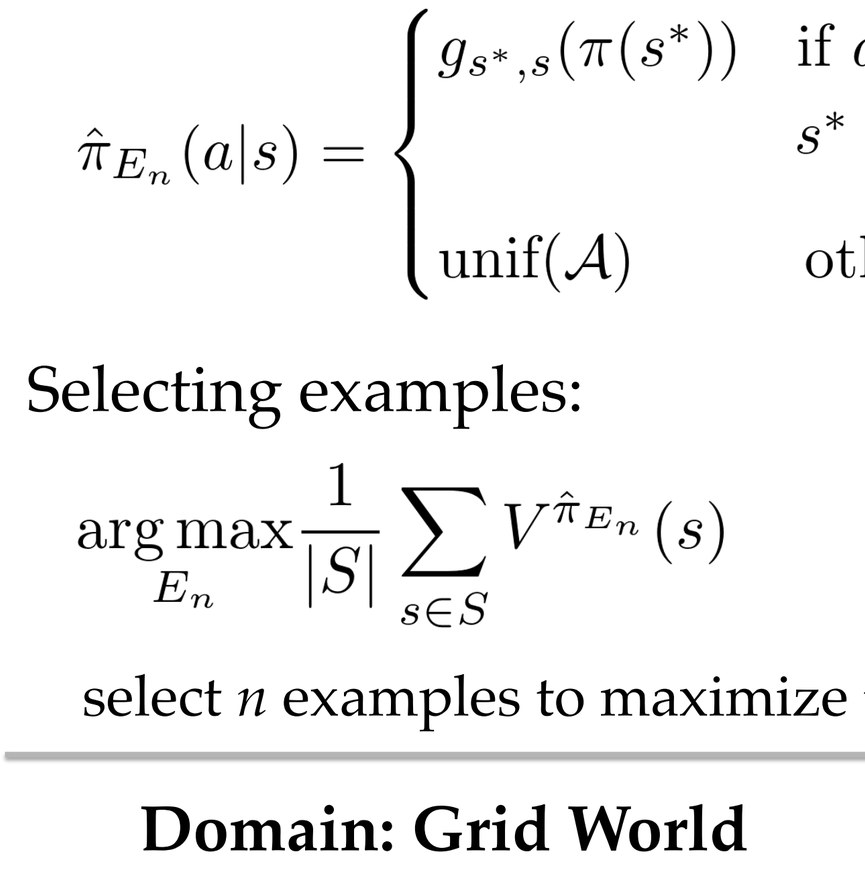
$$\arg\max_{E_n} \frac{1}{|S|}\sum_{s\in S} V^{\hat{\pi}_{E_n}}(s)$$

select n examples to maximize trust

$$\arg\max_{C} \sum_{s\in C}\left[\max_a Q^\pi(s,a) - \frac{1}{|\mathcal{A}|}\sum_a Q^\pi(s,a)\right]$$

greedy approach is optimal, given this human model

Domain: Grid World



Interpretable critical states:
clusters 1, 2, 3 avoid yellow
clusters 4, 5 avoid red

Domain: Atari