Rejecting Hallucinated State Targets during Planning

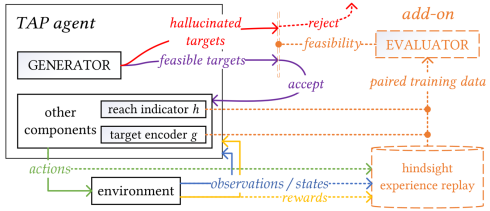
Mingde "Harry" Zhao [1,2,3], Tristan Sylvain [4], Romain Laroche [3], Doina Precup [1,2,5,6] & Yoshua Bengio [1,5,6,7]

1: Mila (Quebec AI Institute), 2: McGill University, 3: Wayve, 4: RBC Borealis, 5: (DeepMind), 6: CIFAR Chair, 7: DIRO (Université de Montréal),

ICML International Conference On Machine Learning

Abstract

Model hallucinations can lead to infeasible targets (sets of states), which cause delusional planning behaviors in Target-Assisted Planning (TAP). This work first categorizes infeasible targets. Then, we propose to reject infeasible targets with an add-on evaluator, which trains alongside TAP agents, requiring almost no change to the agent (and the generative model, "generator") it is assisting. To make sure the learned evaluator is non-delusional, we developed a solution combining 1) specific learning rule, 2) architecture, and 3) assistive hindsight relabeling strategies. Our experiments validate significant reductions in delusional behaviors and performance improvements for several kinds of existing TAP agents.



Targets & τ-feasibility

A target is an embedding of a set of states. Assume we are given h which indicates if a state belongs to target $\boldsymbol{g}^{\odot}$. For practical purposes, we also introduce a time limit τ. Let $D_\pi(s, \boldsymbol{g}^{\odot})$ be a random variable representing the 1st timestep t, s.t. $h(s_t, \boldsymbol{g}^{\odot}) = 1$. We define the τ-feasibility of $\boldsymbol{g}^{\odot}$ from s under π as

$$p\big(D_\pi(s, \boldsymbol{g}^{\odot}) \le \tau\big) \coloneqq \sum_{t=1}^{\tau} p\big(D_\pi(s, \boldsymbol{g}^{\odot}) = t\big)$$

A target is generally considered "good" if it leads to rewarding outcomes:

$$\mathcal{U}_{\pi,\mu}(s, \boldsymbol{g}^{\odot}, \tau) \coloneqq r_\pi(s, \boldsymbol{g}^{\odot}, \tau) + \gamma_\pi(s, \boldsymbol{g}^{\odot}, \tau) \cdot V_\mu\left(s_{\min(D_\pi(s,\boldsymbol{g}^{\odot}),\tau)}\right)$$

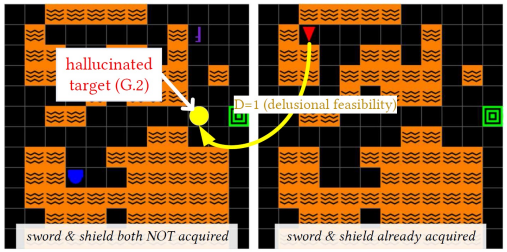
where $\min(D_\pi(s, \boldsymbol{g}^{\odot}), \tau)$ is the timestep the commitment to $\boldsymbol{g}^{\odot}$ is terminated (by h or τ), $s_{\min(D_\pi(s,\boldsymbol{g}^{\odot}),\tau)}$ is the state the agent ended up in, $r_\pi(s, \boldsymbol{g}^{\odot}, \tau) \coloneqq \sum_{t=1}^{\min(D_\pi(s,\boldsymbol{g}^{\odot}),\tau)} \gamma^{t-1} r_t$ is the cumulative discounted reward along the way, $\gamma_\pi(s, \boldsymbol{g}^{\odot}, \tau) \coloneqq \gamma^{\min(D_\pi(s,\boldsymbol{g}^{\odot}),\tau)}$ is the associated cumulative discount, and $V_\mu\left(s_{\min(D_\pi(s,\boldsymbol{g}^{\odot}),\tau)}\right)$ is the future value for following μ from $s_{\min(D_\pi(s,\boldsymbol{g}^{\odot}),\tau)}$.

Infeasible Targets (Singleton Case)

G.1: permanently infeasible targets that do not correspond to any state of the MDP.

G.2: **temporarily** infeasible targets that correspond to a valid state in the MDP, that cannot be reached from the current state. G.2 targets may not exist, depending on the MDP structure.

G.1-induced delusional behavior



G.2-induced delusional behavior (s in $\langle 1, 1\rangle$, $s^{\odot}$ in $\langle 0, 0\rangle$)

Infeasible Targets (Non-singleton Case)

Theorem 4.1: τ-feasibility of a target set is equivalent to the reduced set, *i.e.*, the set with G.1 and G.2 states removed. Infeasible targets: consist of solely G.1 and / or G.2 states

Target Composition	State Correspondence	∞-Feasibility $p(D_\pi(s, g^{\odot}) < \infty)$	Feasibility Delusions & Resulting Delusional Planning Behaviors
Only or Single G.0	non-hallucinated feasible states from s	> 0	**E.0**: May think G.0 states are infeasible, thus turn to riskier alternatives, *e.g.*, G.1 or G.2
Only or Single G.1	hallucinated "states" not belonging to the MDP	should $= 0$	**E.1**: May think G.1 states are favorable, thus commit to them. Impacted by ill-defined $V_\mu(\cdots)$
Only or Single G.2	hallucinated MDP states infeasible from s	should $= 0$	**E.2**: May think G.2 states are favorable, thus commit to them
Some G.0	at least one non-hallucinated state from s	$= p(D_\pi(s, g^{\odot}) < \infty) > 0$ (Thm. 4.1)	**E.0**
Only G.1 & G.2	set of ONLY hallucinated states	should $= 0$	**E.1** & **E.2**

SOLUTION DESIGN: An auxiliary evaluator that can be used to reject infeasible targets

An evaluator that can estimate the τ-feasibility of targets with the input of source-target pairs $\langle s, g^{\odot}\rangle$.

[automatic] the evaluator learns to automatically differentiate the feasibility of all kinds of targets without pre-labeling: we need to exploit h

[minimally intrusive] the evaluator is made to be generally applicable to existing TAP agents without changing the agents too much to disturb the sensitive RL components: we need to ensure its behavior as an add-on and it can be conditioned on the policy π of the agent, to learn alongside the agent.

[unified] the evaluator has a unified behavior compatible with different τs: we design it in a way to learn the τ-feasibilities for many τs simultaneously.

SOLUTION IMPLEMENTATION: combo of 3

1) **Learning Rule (indirect)**
2) **Distributional Architecture**
3) **Training Data (Source-Target Pair Construction & hindsight-relabeling)**

Evaluator Learning Rule

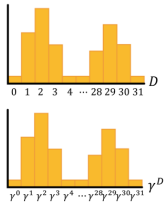
Indirectly learn feasibility by learning the distribution of $D_\pi(s, \boldsymbol{g}^{\odot})$. Represent $D_\pi(s, \boldsymbol{g}^{\odot})$ as $D_\pi(s, a, \boldsymbol{g}^{\odot})$, where $a \sim \pi(\cdot\,|s, \boldsymbol{g}^{\odot})$

$D_\pi(s, \boldsymbol{g}^{\odot}) \leftarrow 1 + D_\pi(s', \boldsymbol{g}^{\odot})$, with

$$\begin{cases} D_\pi(s', \boldsymbol{g}^{\odot}) \coloneqq \infty & \text{if } s' \text{ is terminal and } h(s', \boldsymbol{g}^{\odot}) = 0 \\ D_\pi(s', \boldsymbol{g}^{\odot}) \coloneqq 0 & h(s', \boldsymbol{g}^{\odot}) = 1 \end{cases}$$

Evaluator Architecture

C51-distributional output with each bin probability being $p(D_\pi(s, \boldsymbol{g}^{\odot}) = t)$. Compatible with varying time-horizons Swap support for different outputs

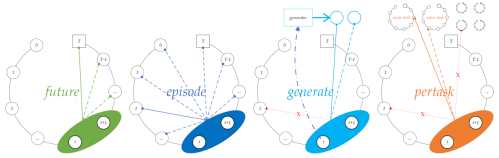


Evaluator Training Data

The evaluator evaluates and trains on source-target pairs $\langle s, \boldsymbol{g}^{\odot}\rangle$, these pairs can be constructed non-trivially with hindsight-relabeling.

Training-Reasoning Discrepancy

most existing target-directed agents are only trained on experienced data and thus only know how to deal with feasible targets. Yet, they must also deal with hallucinated targets during planning. **Naïve relabeling will cause delusions (false evaluations that cannot be fixed by more training)!**

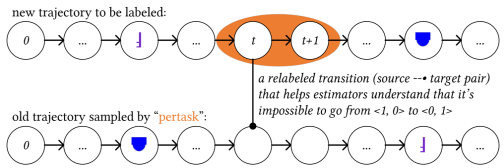


"generate" strategy: Let Evaluator Learn about Generated Candidates

expose targets that the generator can generate

"pertask" strategy: Let Evaluator Learn about Experienced Targets Outside the Episode

expose targets previously achieved in other episodes



Experiment Summary

Evaluator can help reduce delusional planning behaviors and boost performance of different types of TAP agents