# Incoherence in goal-conditioned autoregressive models

Jacek Karwowski  Raymond Douglas

## 1. What is incoherence?

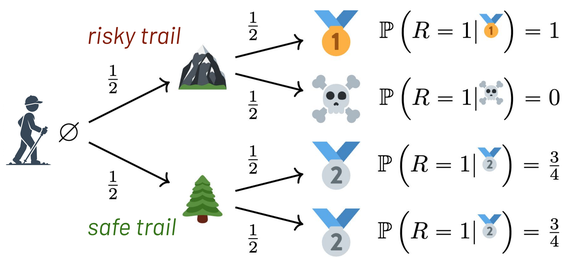
*In an MDP with a prior over actions $p(a|s)$ and a binary reward $R$, the goal-conditioned policy $\pi(a|s) = p(a|s, R = 1)$ answers the question:*

? Which action to take in state $s$, such that, if later choices are sampled from the prior $p(a|s)$ the outcome will lead to $R$?

*rather than*

? Which action to take in state $s$, such that, if later choices are sampled from the policy $\pi(a|s)$ the outcome will lead to $R$?

## 2. Example: mountain race




The goal-conditioned policy is computed as:

$$\pi(a|\varnothing) = p\left(a|\varnothing, R = 1\right) = \frac{p\left(R = 1|\varnothing, a\right) p\left(a|\varnothing\right)}{p\left(R = 1|\varnothing\right)} \propto p\left(R = 1|\varnothing, a\right)$$

evaluating later actions according to the prior:

$$\pi(\nearrow|\varnothing) \propto p\left(R = 1|\varnothing, \nearrow\right) = p\left(R = 1|\text{mountain}\right) = \frac{1}{2}$$

$$\pi(\searrow|\varnothing) \propto p\left(R = 1|\varnothing, \searrow\right) = p\left(R = 1|\text{tree}\right) = \frac{3}{4}$$

even though the goal-conditioned policy would avoid the chasm!

## 3. Formalising incoherence

Incoherence is the gap between $\pi$ and the policy recommended by $\pi$'s own future Q-values:

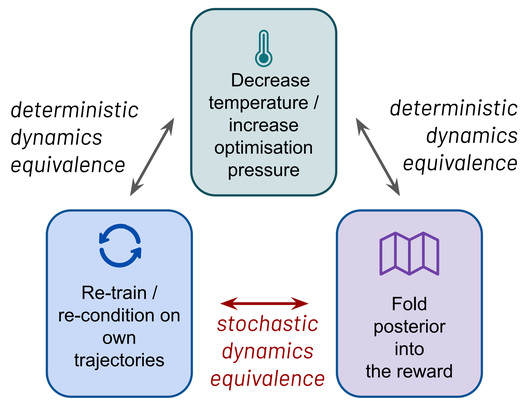
$$\text{Incoherence}_\tau(\pi) = \text{KL}_\xi\left[\pi(\xi) \;||\; \text{Boltzmann}_\tau(Q^\pi)(\xi)\right]$$

over trajectories $\quad \xi = (s_0, a_0, \ldots s_n)$

where $\quad \text{Boltzmann}_\tau(Q^\pi)(a|s) \;\propto\; \exp\left(\frac{Q^\pi(a,s)}{\tau}\right)$

## 4. Three-way correspondence

Three equivalent views of policy tightening:




**Theorem** (Optimisation pressure, thrice). *For all priors $p(a|s)$, all non-positive reward functions $r(s_t, a_t)$ and deterministic transition dynamics $\tau(s_{t+1}|s_t, a_t)$, for all $k \in \mathbb{N}_+$ and for $\alpha(k) = 2^k$, we have:*

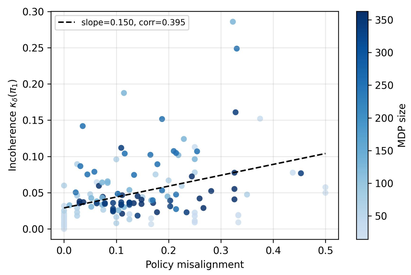
$$\pi_{\alpha(k)} = \pi^{\mathcal{G}}_{2^k} = \pi^{\mathcal{F}}_{2^k} = \pi^{\mathcal{H}}_k$$

*Additionally, for stochastic transition dynamics we have:*

$$\pi^{\mathcal{F}}_k = \pi^{\mathcal{G}}_k$$

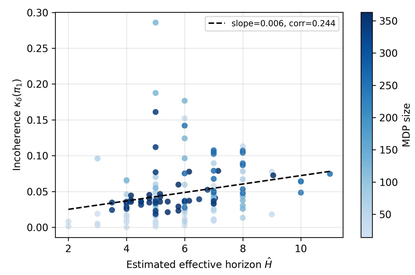
## 5. Effective horizon

*When actions with the highest Q-values under the random policy also have the highest Q-values under the optimal policy (i.e. when it is optimal to be greedy on the random policy's Q-function), deep RL tends to succeed; when they don't, deep RL tends to fail. (Laidlaw et al., 2023)*




Policy misalignment: the proportion of actions being different between optimal policy and a policy selecting the greedy action under the random policy's Q-function (averaged over the latter's occupancy measure)

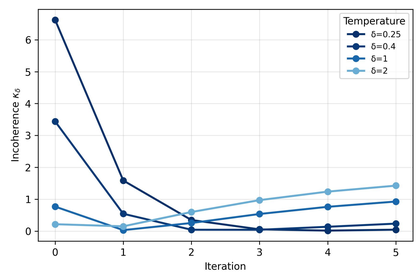
Monte Carlo effective horizon estimation:

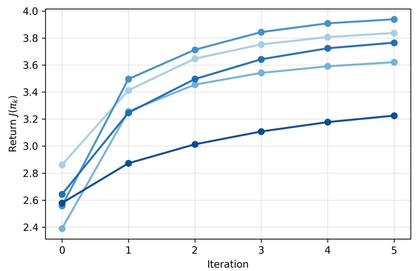



## 6. Incoherence as a metric

A corollary of the theorem:

$$\lim_{(\tau,i)\to(0,\infty)} \text{Incoherence}_\tau(\pi^{\mathcal{G}}_i) = 0$$




The average correlation between return and negative incoherence across a dataset of trajectories: 0.863

github.com/jkarwowski/incoherence  arXiv:2510.06545 →