Empowerment, Free Energy Principle and Maximum Occupancy Principle (MOP) Compared

MOP paper v1

Rubén Moreno-Bote (1,2), Demetrio Ferro (1), and Jorge Ramírez-Ruiz (1)

(1) Center for Brain and Cognition, and Department of Information and Communication Technologies, Universitat Pompeu Fabra, Barcelona, Spain
(2) Serra Húnter Fellow Programme, Universitat Pompeu Fabra, Barcelona, Spain

Motivation

Natural behavior, even stereotyped one, **is variable**. The reasons for this variability are unknown. We **propose** that the **goal of behavior is to produce guided variability**, i.e., generate all sorts of action-state paths compatible with the dynamics and constraints of the agent. We call this **Maximum Occupancy Principle (MOP)**. We compare MOP with other two reward-free approaches in Markov Decision Processes (MDP): Empowerment and Free Energy Principle

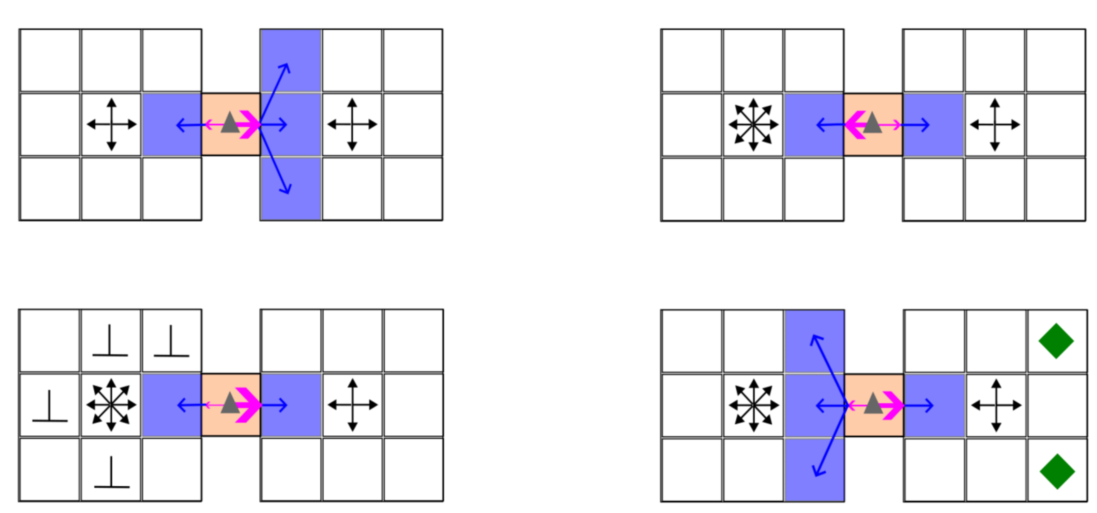
Maximum Occupancy Principle (MOP)

Goal: Maximize future cumulative action-state path entropy [1]
Bias: Agents prefer states that promise future action-state entropy (freedom & exploration) while avoiding absorbing states (survival instinct)
Recursive?: *Yes*, a Bellman equation can be written

$\pi(a_t|s_t)$ policy — $p(s_{t+1}|s_t,a_t)$ state transition prob. — $\tau \equiv (s_0,a_0,s_1,...,a_t,s_{t+1},...)$ action-state path (trajectory)

return
$$R(\tau) = -\sum_{t=0}^{\infty}\gamma^t \ln\left(\pi^\alpha(a_t|s_t)p^\beta(s_{t+1}|s_t,a_t)\right)$$

value
$$V_\pi(s) \equiv \mathbb{E}_{a_t\sim\pi(a_t|s_t),s_{t+1}\sim p(s_{t+1}|s_t,a_t)}[R(\tau)|s_0=s]$$
$$= \mathbb{E}_{a_t\sim\pi(a_t|s_t),s_{t+1}\sim p(s_{t+1}|s_t,a_t)}\left[\sum_{t=0}^{\infty}\gamma^t\left(\alpha\mathcal{H}(A|s_t)+\beta\mathcal{H}(S'|s_t,a_t)\right)\Big|s_0=s\right]$$



$$V^*(s) = \sum_{s',a}\pi^*(a|s)p(s'|s,a)\Big[\underbrace{-\ln\left(\pi^{*\alpha}(a|s)p^\beta(s'|s,a)\right)}_{\text{Immediate occupancy}} + \underbrace{\gamma V^*(s')}_{\text{Future occupancy}}\Big]$$

→ available actions in room ⊥ absorbing state ◆ food

Empowerment (MPOW)

Goal: Maximize mutual information between a sequence of actions and the resulting state [2]
Bias: Agents prefer empowered states, i.e., unstable fixed points of the dynamics
Recursive?: *No*, a Bellman equation cannot be written, because Mutual Info is not additive [1], but approximations exist [1,3]

$a_t^n = (a_t,a_{t+1},...,a_{t+n-1}) \in \mathcal{A}^n$ planned sequence of actions — $p(s_{t+n}|s_t,a_t^n) = \tau(a_t^n|s_t)\prod_{\tau=t}^{t+n-1}p(s_{\tau+1}|s_\tau,a_\tau)$ n-step transition probability

$$\mathcal{E}(s_t) = \max_{\tau(a_t^n|s_t)}\sum_{a_t^n,s_{t+1}}p(s_{t+n}|s_t,a_t^n)\tau(a_t^n|s_t)\log\left(\frac{p(s_{t+n}|s_t,a_t^n)}{\sum_{a_t^n}p(s_{t+n}|s_t,a_t^n)\tau(a_t^n|s_t)}\right)$$

state empowerment; transitions are greedy towards the accessible state with highest empowerment

Free Energy Principle (FEP / EFE)

Goal: Minimize KL divergence between actual and target distributions [4]
Bias: Agents prefer states where target distribution peaks (preferred states), and behavior tends to collapse to a deterministic policy around them
Recursive?: *Yes* in fully observable MDPs ('sophisticated inference')

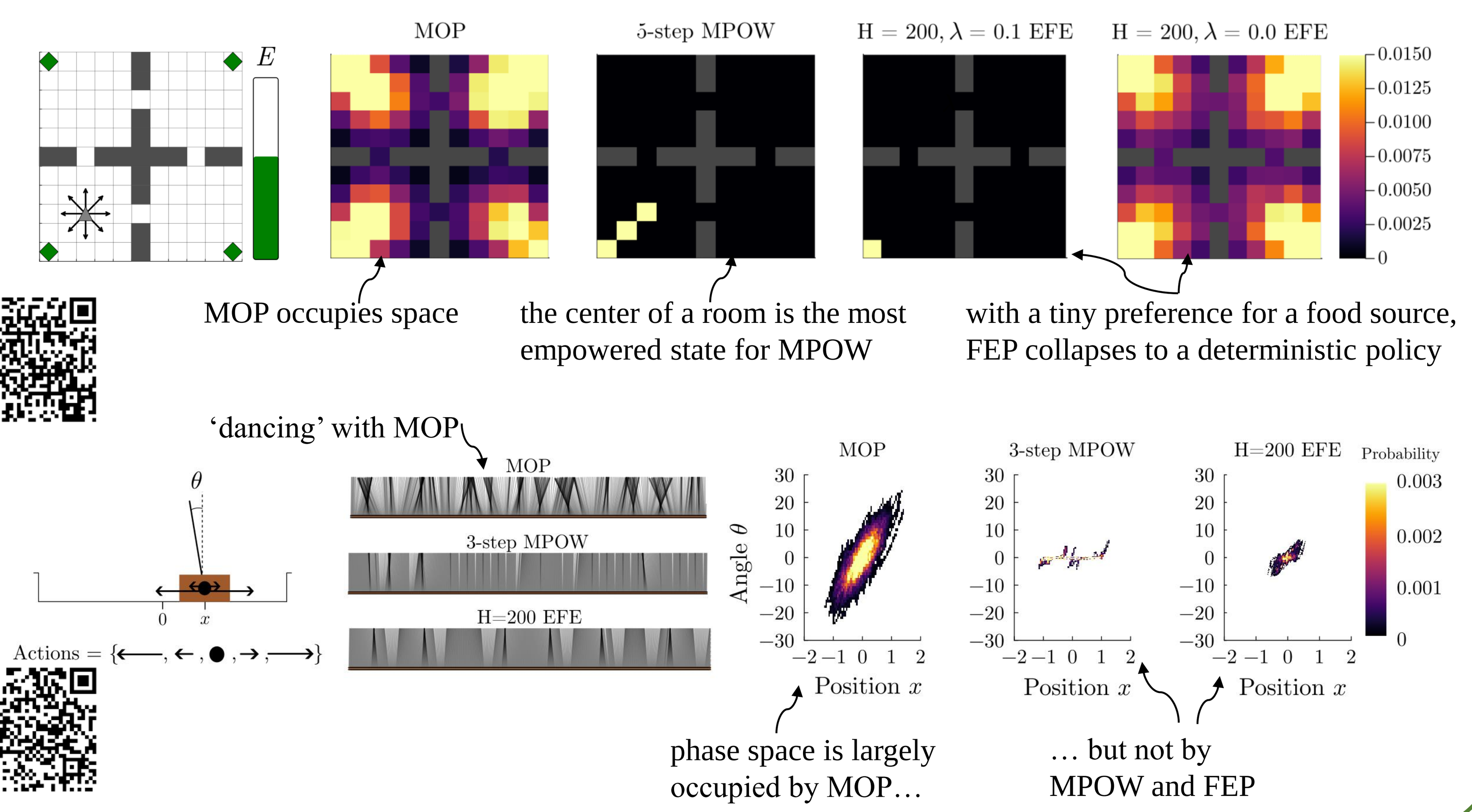
$a_t^{T-1} = (a_t,a_{t+1},...,a_{T-1})$ action path up to a horizon — $s_{t+1}^T = (s_{t+1},s_{t+2},...,s_T)$ state path up to a horizon — $q(s_{t+1}^T) = \prod_{\tau=t}^{T-1}q(s_{\tau+1})$ target distribution factorizes

cost
$$G_{\pi,t}(s_t) = \mathbb{E}_{a_t^{T-1}\sim\pi}\mathrm{KL}\left(p(s_{t+1}^T|a_t^{T-1},s_t)||q(s_{t+1}^T)\right)$$
$$= \sum_{s_{t+1}^T,a_t^{T-1}}p_\pi(s_{t+1}^T,a_t^{T-1}|s_t)\log\frac{p(s_{t+1}^T|a_t^{T-1},s_t)}{q(s_{t+1}^T)}$$

Bellman eq.
$$G_{\pi,t}(s_t) = \sum_{s_{t+1},a_t}\pi(a_t|s_t)p(s_{t+1}|s_t,a_t)\left[\log\frac{p(s_{t+1}|s_t,a_t)}{q(s_{t+1})}+G_{\pi,t+1}(s_{t+1})\right]$$

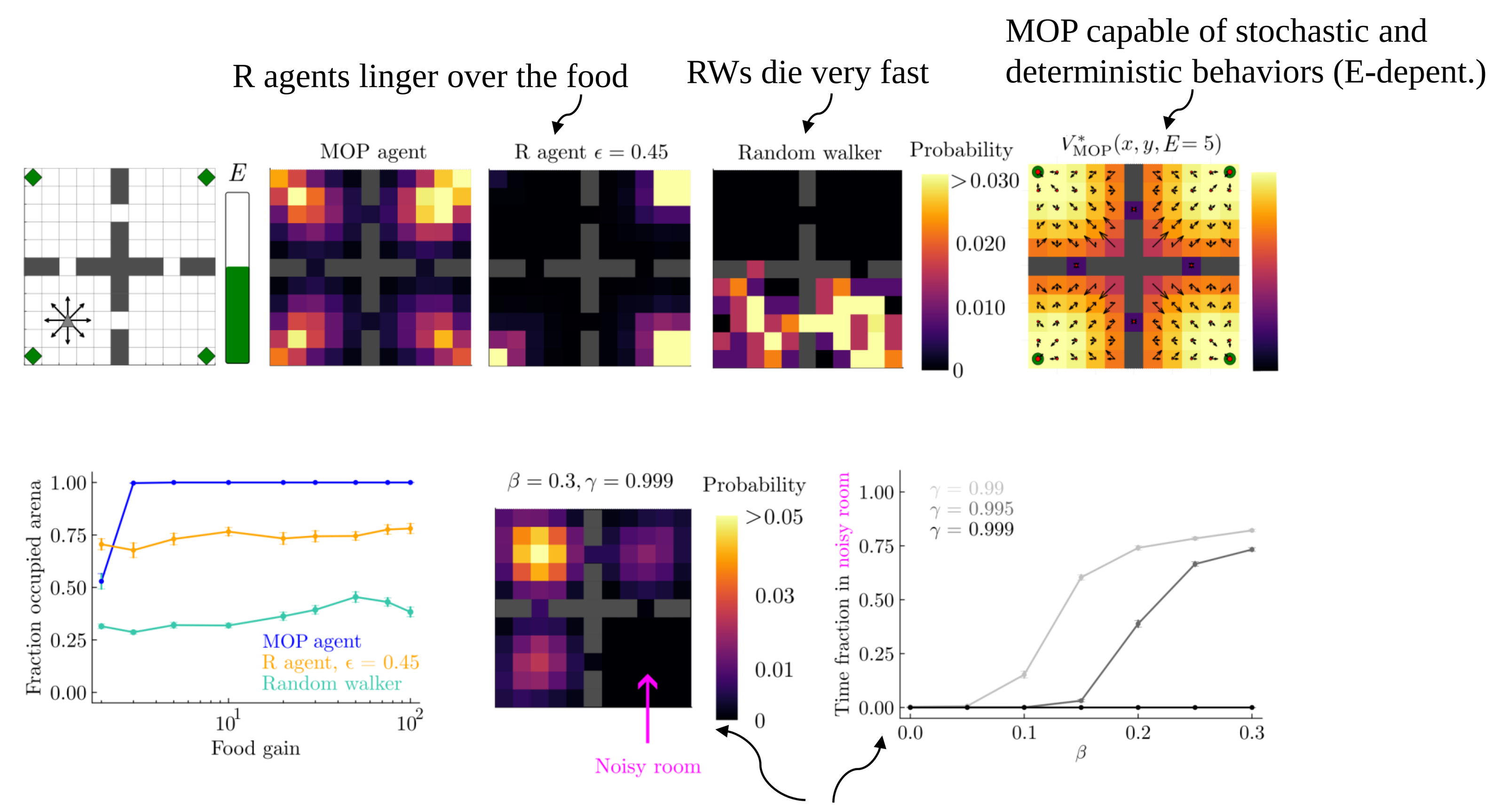
Main Result: MOP, MPOW & FEP compared

MOP produces action-state entropy non-stop, while MPOW favors unstable fix points and FEP collapses to deterministic behaviors in fully observable MDPs. This is observed in two very common environments, a grid-world and a cartpole: use QRs for compelling examples



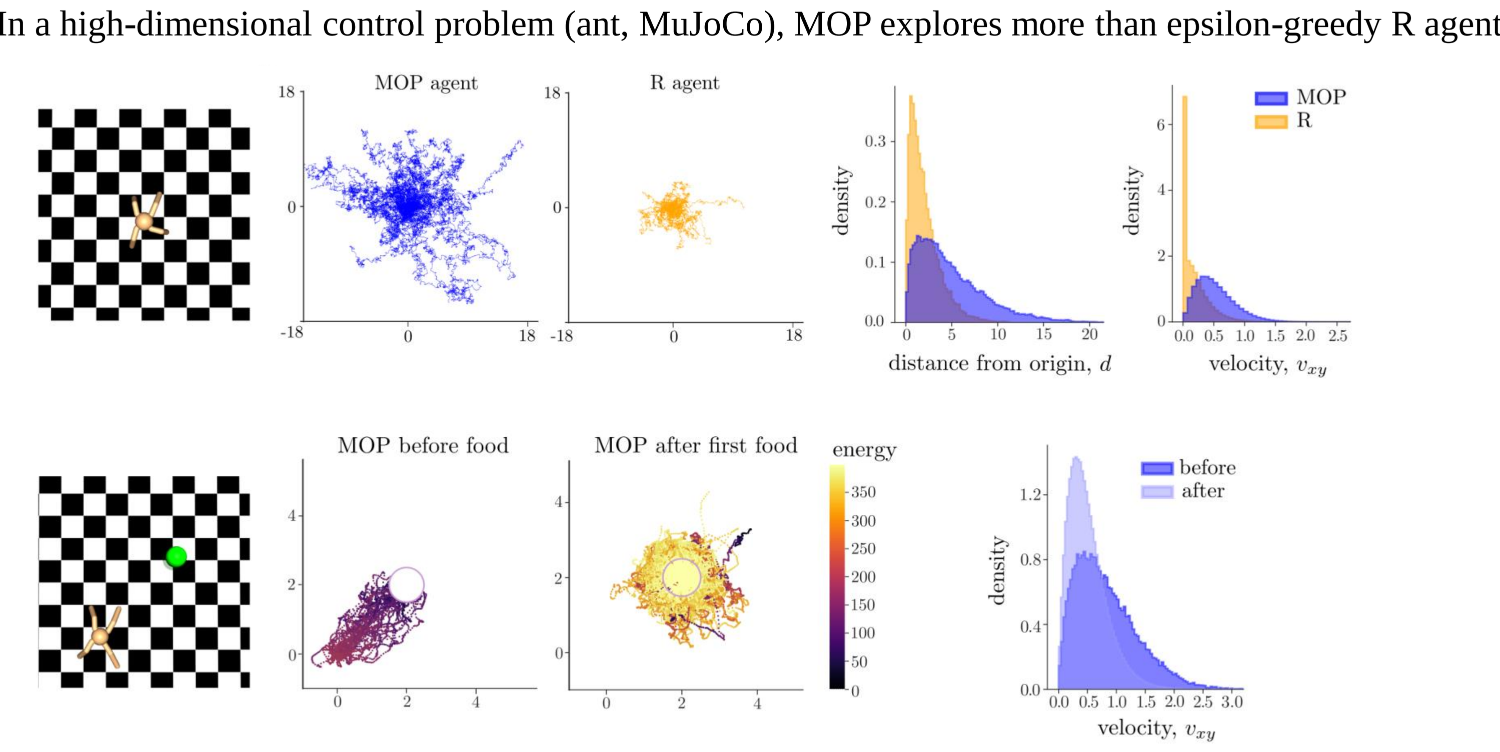
MOP vs Reward Maximization

MOP generates complex behavior in both a grid-world and ant environment.
In contrast, epsilon-greedy reward maximization matching survival times leads to less variable behaviors



MOP avoids a "noisy TV room" if gamma is large enough

In a high-dimensional control problem (ant, MuJoCo), MOP explores more than epsilon-greedy R agents



MOP develops both stochastic and deterministic state-dependent policies

References

[1] Jorge Ramírez-Ruiz, Dmytro Grytskyy, and Rubén Moreno-Bote. Seeking entropy: com plex behavior from intrinsic motivation to occupy action-state path space. arXiv preprint arXiv:2205.10316, 2022.
[2] Tobias Jung, Daniel Polani, and Peter Stone. Empowerment for continuous agent—environment systems. Adaptive Behavior, 19(1):16–39, 2011.
[3] Felix Leibfried, Sergio Pascual-Diaz, and Jordi Grau-Moya. A unified bellman optimality principle combining reward maximization and empowerment. Advances in Neural InformationProcessing Systems, 32, 2019.
[4] Lancelot Da Costa, Noor Sajid, Thomas Parr, Karl Friston, and Ryan Smith. Reward maximization through discrete active inference. Neural Computation, 35(5):807–852, 2023.

Funding