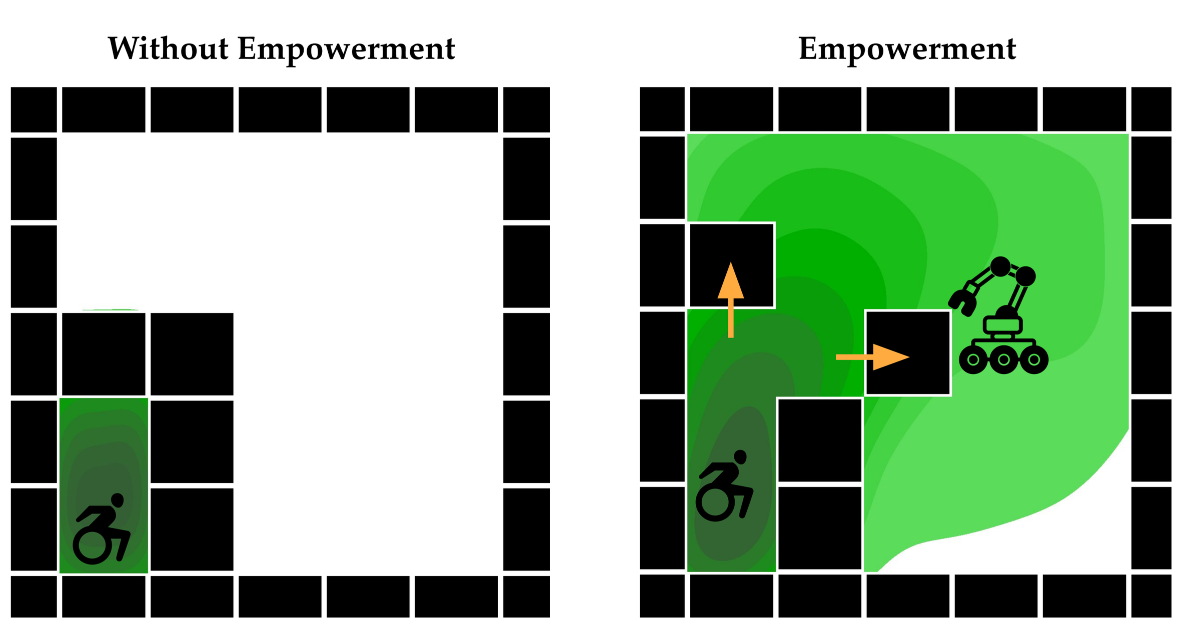
Learning to Assist Humans without Inferring Rewards

Vivek Myers, Evan Ellis, Sergey Levine, Benjamin Eysenbach, Anca Dragan

RAIL · InterACT Laboratory · BAIR Berkeley Artificial Intelligence Research · Princeton University

Motivation

- Assistive agents typically assume humans are optimizing a reward (e.g., CIRL [1] setup)
  - Learning rewards is hard
  - Misspecified rewards can be very harmful
  - Human objectives often not well-modeled by rewards
- How can we create aligned agents without learning and maximizing a human reward?
- **Key Idea:** maximize an **empowerment objective** [2,3,4,5,6] to help the human have maximal control over the world
  - We show how **a scalable contrastive algorithm can estimate and maximize human empowerment**



Notation

- Two agents interact in an MDP

$$M = (\mathcal{S}, \mathcal{A}_{\mathbf{H}}, \mathcal{A}_{\mathbf{R}}, R, P, \gamma)$$

  - Human **H**
  - Assistive agent ("robot") **R**
- Two policies $\mathfrak{a}_t^{\mathbf{R}} \sim \pi_R(\bullet \mid \mathfrak{s}_t)$ and $\mathfrak{a}_t^{\mathbf{H}} \sim \pi_H(\bullet \mid \mathfrak{s}_t)$
- Dynamics $P(s' \mid s, a^{\mathbf{H}}, a^{\mathbf{R}})$
  - random variables $\mathfrak{s}_t$ represent state at time t
  - future state $\mathfrak{s}^+$ is a random $\mathfrak{s}_K$ with

$K \sim \mathrm{Geom}(1-\gamma)$

Objective: Empowering Humans

- For policies π_H and π_R we define the **human empowerment** objective:

$$\mathcal{E}(\pi_{\mathbf{H}}, \pi_{\mathbf{R}}) = \mathbb{E}\Big[\sum_{t=0}^{\infty} \gamma^t I(\mathfrak{a}_t^{\mathbf{H}}; \mathfrak{s}^+ \mid \mathfrak{s}_t)\Big]$$

Empowerment

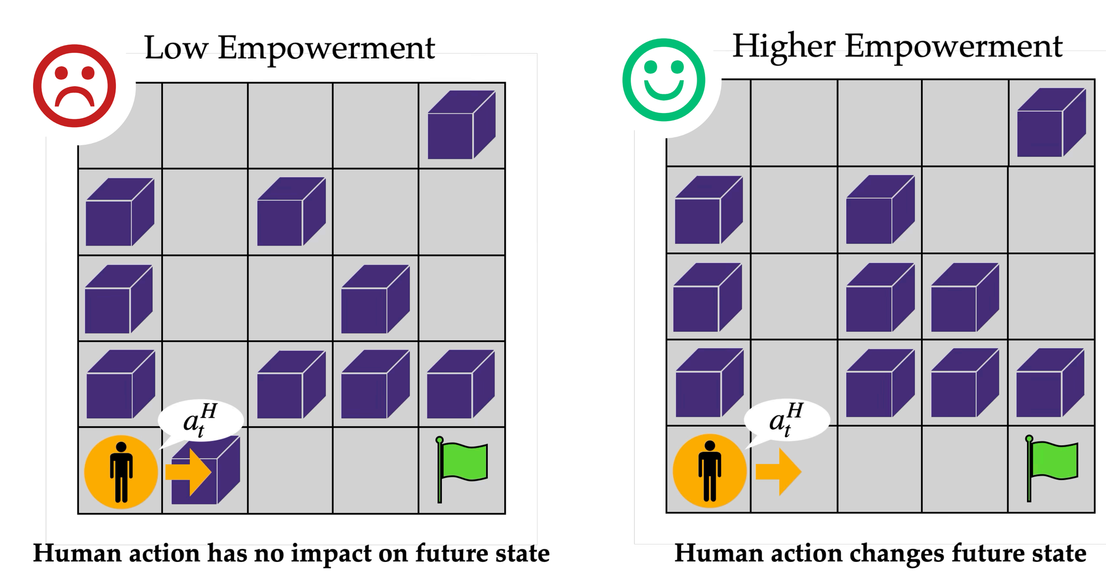
Sum over future steps

Mutual Informations between human action and the future state

- The mutual information term becomes:

$$I(a_t^{\mathbf{H}}; \mathfrak{s}^+ \mid s_t) \triangleq \mathbb{E}_{s_t, s_{t+k}, a_t^{\mathbf{H}}, a_t^{\mathbf{R}}}\left[\log \frac{p(\mathfrak{s}_{t+K} = s_{t+k} \mid \mathfrak{s}_t = s_t, \mathfrak{a}_t^{\mathbf{H}} = a_t)}{p(\mathfrak{s}_{t+K} = s_{t+k} \mid \mathfrak{s}_t = s_t)}\right]$$

When $\mathcal{E}(\pi_H, \pi_R)$ is large, the human's actions have high influence on the (γ-discounted) future state



Analysis

What does empowerment do when the human *is* optimizing a scalar reward?

Theorem 3.1. *Under Assumption 3.1 and Assumption 3.2, for sufficiently large γ and any $\beta > 0$,*

$$\mathcal{E}_\gamma(\pi_{\mathbf{H}}, \pi_{\mathbf{R}})^{1/2} \le (\beta/e)\, \mathcal{J}_{\pi_{\mathbf{R}}}^{\gamma}(\pi_{\mathbf{H}}).$$

Robot empowerment objective

Discounted return under human's objective

Assumption 3.1 (Skill Coverage). *The rewards $R \sim \mathcal{R}$ are uniformly distributed over the scaled $|\mathcal{S}|$-simplex $\Delta^{|\mathcal{S}|}$ such that:*

$$\left(R + \tfrac{1}{|\mathcal{S}|}\right)\left(\tfrac{1}{1-\gamma}\right) \sim \mathrm{Unif}\left(\Delta^{|\mathcal{S}|}\right) = \mathrm{Dirichlet}(\underbrace{1, 1, \dots, 1}_{|\mathcal{S}| \text{ times}}).$$

Assumption 3.2 (Ergodicity). *For some $\pi_{\mathbf{H}}, \pi_{\mathbf{R}}$, we have*

$$\mathrm{P}^{\pi_{\mathbf{H}}, \pi_{\mathbf{R}}}(\mathfrak{s}^+ = s \mid s_0) > 0 \quad \text{for all } s \in \mathcal{S}, \gamma \in (0, 1).$$

- Increasing human empowerment optimizes a **lower bound** on the **average-case reward**

References

[1] Hadfield-Menell, D. et al., 2016. "Cooperative Inverse Reinforcement Learning." *NeurIPS*
[2] Du, Y. et al., 2020. "AvE: Assistance via Empowerment." *NeurIPS*
[3] Salge, C. et al., 2014. "Empowerment–an Introduction." *Guided Self Organization Inception*
[4] Choi, J. et al., 2021. "Variational Empowerment as Representation Learning for Goal-Conditioned Reinforcement Learning." *ICML*
[5] Mohamed, S. et al., 2015. "Variational Information Maximisation for Intrinsically Motivated Reinforcement Learning." *NeurIPS*
[6] Bharadhwaj, H. et al., 2022. "Information Prioritization Through Empowerment in Visual Model-Based Rl." *ICLR*
[7] Oord, A. et al., 2018. "Representation Learning With Contrastive Predictive Coding." *arXiv*
[8] Poole, B. et al., 2019. "On Variational Bounds of Mutual Information." *ICML*
[9] Carroll, M. et al., 2019. "On the Utility of Learning About Humans for Human-AI Coordination." *NeurIPS*

Learning Empowerment

- We use a **time-contrastive objective** [7,8] to estimate empowerment:

$$\mathcal{L}_c(\{x_i\}, \{y_j\}) \triangleq \sum_{i=1}^{N}\Bigg[\log\Bigg(\overbrace{\frac{e^{x_i^T y_i}}{\sum_{j=1}^{N} e^{x_i^T y_j}}}^{\textit{predict future from present}}\Bigg) + \log\Bigg(\overbrace{\frac{e^{x_i^T y_i}}{\sum_{j=1}^{N} e^{x_j^T y_i}}}^{\textit{predict past from future}}\Bigg)\Bigg]$$

$$\max_{\phi, \phi', \psi} \mathbb{E}_{\{(s_i, a_i, s_i') \sim \mathrm{P}(\mathfrak{s}_t, \mathfrak{a}_t^{\mathbf{H}}, \mathfrak{s}_{t+k})\}_{i=1}^{N}}\Big[\overbrace{\mathcal{L}_c(\{\phi(s_i, a_i)\}, \{\psi(s_j')\})}^{\textit{conditioned on human action}} + \overbrace{\mathcal{L}_c(\{\phi'(s_i)\}, \{\psi(s_j')\})}^{\textit{unconditional classification}}\Big]$$

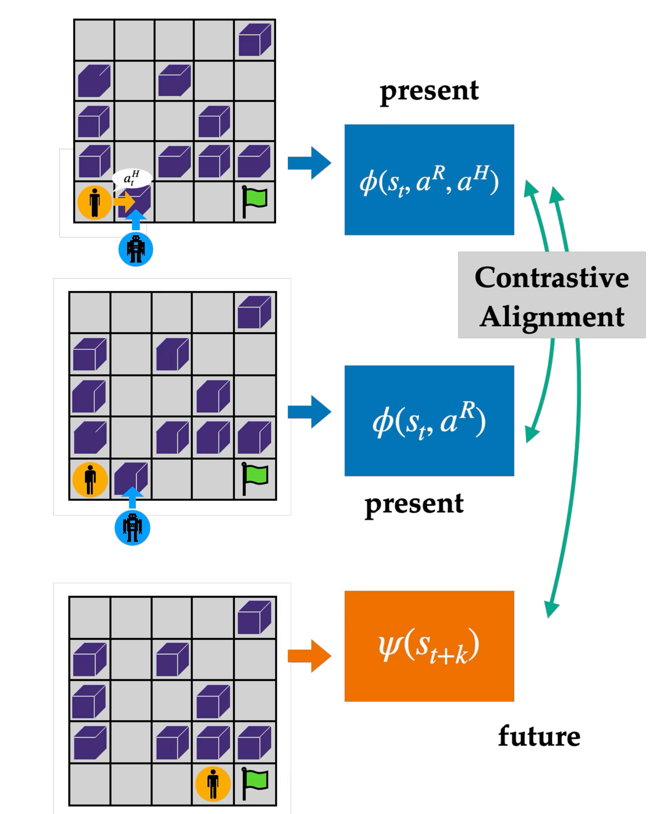
- Get probability ratios both conditioned and unconditioned on the human action $a^{\mathbf{H}}$

  …which lets us approximate empowerment:

$$\mathcal{E}(\pi_{\mathbf{H}}, \pi_{\mathbf{R}}) \approx \mathbb{E}_{\pi_{\mathbf{H}}, \pi_{\mathbf{R}}}\Big[\sum_{t=0}^{\infty} \gamma^t (\phi(s_t, a^{\mathbf{R}}, a^{\mathbf{H}}) - \phi(s_t, a^{\mathbf{R}}))^T \psi(g) - \log \tfrac{C_2}{C_1}\Big].$$

- Use this estimate as the **proxy reward** for training the robot with RL:

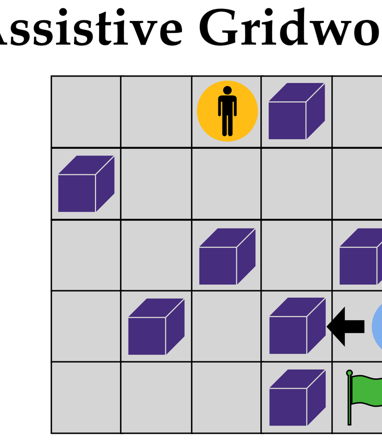
$$\boxed{r(s, a^{\mathbf{R}}) = (\phi(s_t, a^{\mathbf{R}}, a^{\mathbf{H}}) - \phi(s_t, a^{\mathbf{R}}))^T \psi(g).}$$



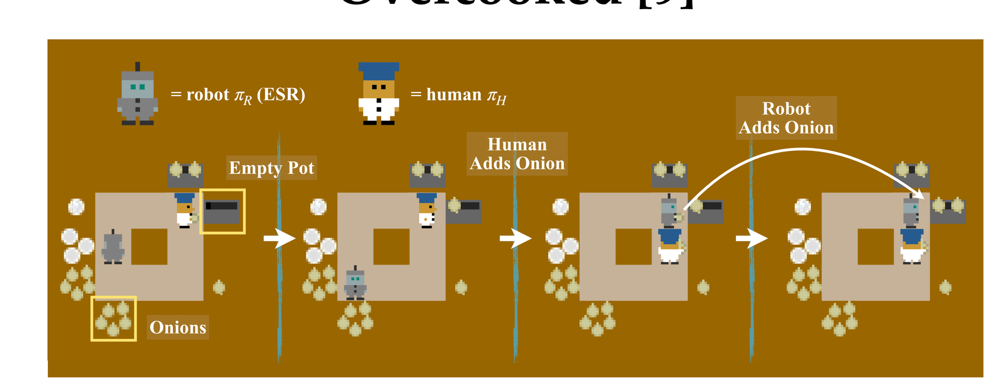
Experiments

Train our approach with a human model and no knowledge of the true human objective

Assistive Gridworld [2]



Overcooked [9]



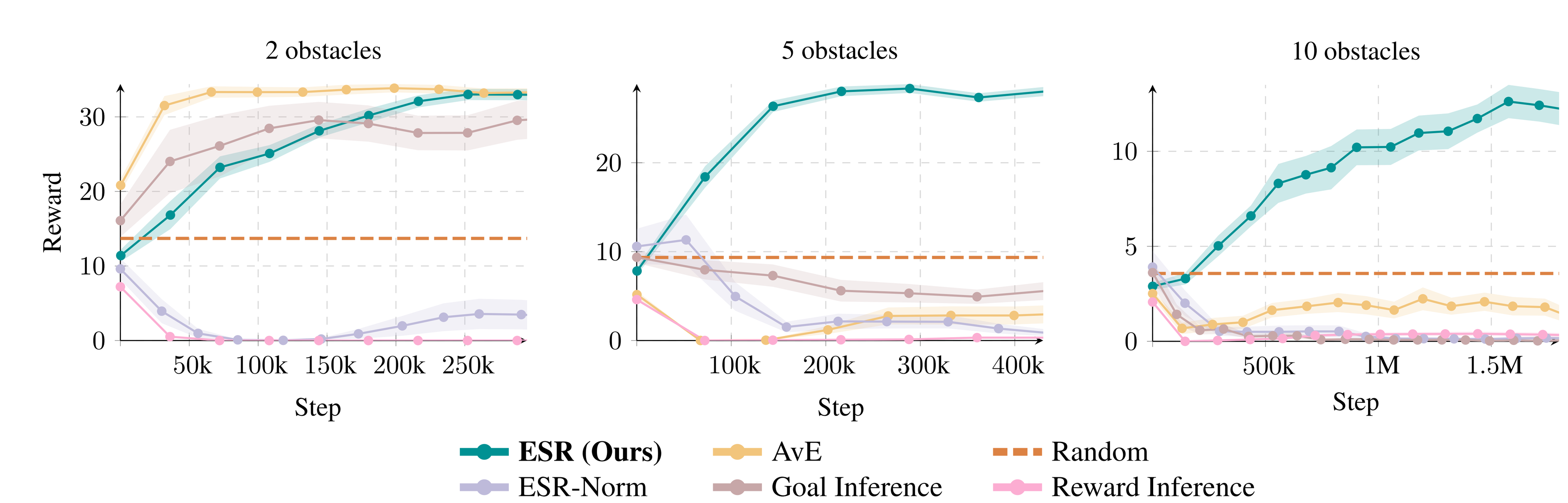


Table 1: Overcooked Results

Layout	**ESR (Ours)**	Reward Inference	AvE	Random
Asymmetric Advantages	$\mathbf{72.00 \pm 5.37}$	60.33 ± 0.26	36.71 ± 1.71	59.36
Coordination Ring	$\mathbf{8.40 \pm 0.69}$	5.96 ± 0.20	5.69 ± 0.93	6.02
Cramped Room	$\mathbf{91.33 \pm 4.08}$	39.24 ± 0.35	5.13 ± 1.31	69.26